METHODS FOR MULTIVARIATE LONGITUDINAL COUNT AND DURATION MODELS WITH APPLICATIONS IN ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Darcy Steeg Morris

August 2012

METHODS FOR MULTIVARIATE LONGITUDINAL COUNT AND DURATION MODELS WITH APPLICATIONS IN ECONOMICS

Darcy Steeg Morris, Ph.D.

Cornell University 2012

Quality and quantity of social science data is continually improving, from large public-use survey microdata to private industry data. This wealth of data allows researchers to ask more complex questions about interdependencies of social and economic processes and behavior. This dissertation presents methods for models that address interdisciplinary research questions about the association structure of multiple outcomes of similar or disparate types, e.g. count and duration outcomes. The proposed models and methods address associations of multiple outcomes through correlated unobserved subject-specific effects.

Chapter 2 presents a semiparametric method for estimating the marginal response and association parameters in a random effects multivariate longitudinal count model. In the context of the generalized estimating equations (GEE) framework, we use a specific form of the covariance matrix of the response vector based on a model that induces dependence over time and outcomes using random effects. This moment based method is robust to distributional misspecification and reduces the computational burden associated with a high-dimensional joint distribution by avoiding parametric assumptions on the response and unobserved effects. Through a simulation study we compare finite sample robustness properties of this semiparametric method with a pseudo-likelihood approach that imposes distributional assumptions. Both of these methods are then used to analyze a

dataset of insurance claim counts for three types of coverage over time. The economic significance of these results is presented in Chapter 3.

Chapter 4 presents a Gaussian variational approximation (GVA) approach for estimation of a joint multivariate longitudinal count and multivariate duration random effects model. GVA proposes an approximate posterior distribution of the random effects to obtain a closed form lower bound of the marginal likelihood. GVA estimators are obtained by maximizing the variational lower bound, which coincides with minimizing the Kullback-Leibler distance between the random effects posterior distribution and the assumed approximate posterior distribution. This approach circumvents the computationally complex, high-dimensional integral associated with the marginal distribution of a joint longitudinal and duration model. Through a simulation study we compare finite sample properties of the variational approximation approach with comparable univariate and multivariate two-stage plug-in approaches. These methods are then used to analyze a dataset of insurance claim counts and policy duration for three types of coverage over time.

BIOGRAPHICAL SKETCH

Darcy Steeg Morris was born and raised in Ramsey, NJ. She completed her undergraduate studies at Princeton University in 2003 with a major in Economics and a certificate in Finance. She spent four post-baccalaureate years working in economic consulting and pursuing a Master's degree in Statistics at The George Washington University in Washington, DC. She began her PhD work at Cornell University in the Department of Statistical Science in the Fall of 2007. Following her doctorate, Darcy will join the "Missing Data" research group as a Research Mathematical Statistician in the Center for Statistical Research and Methodology at the U.S. Census Bureau.

To Brent: Thank You.

To Grandpa Steeg: 60 Years Later.

ACKNOWLEDGEMENTS

Thank you to my committee: Francesca Molinari, Rob Strawderman and Jim Booth. Special thanks to my advisor, Francesca Molinari, for her genuine support and invaluable guidance. You are a true mentor and an inspiring influence both professionally and personally. Thank you to Levon Barseghyan for his enthusiastic interest and determination in finding the truth that the data reveals. You have been key to my understanding of the economic implications of this research. Francesca and Levon, there are no words to fully express my gratitude. Thank you. Thanks to Josh Teitelbaum for securing such a unique and interesting dataset, and for his significant contributions to this work.

I would have never accomplished what I have without my husband's love, humor and support. Brent, thank you for sacrificing for my academic pursuits and for your ever-constant positive outlook on life. You truly make me so happy. Thanks to my family, Mom for her excitement in my successes and unwavering support in my failures; Dad for always being a proud father; Bryce for little pieces of advice that put me and kept me on this path; and Maddy and Kendall for always making me smile.

I am fortunate to have spent my doctoral studies with a supportive, smart and fun set of classmates. Special thanks to Caitlin, Kirsten and Raj. Thank you to my colleagues at CISER for a great four years that has kept my statistical software knowledge current and strong. I will fondly look back on my experience at Cornell, grateful to have been part of such a friendly and intelligent community.

Part of this dissertation was written under the support of the NSF through Grant SES-1031136 to Barseghyan, Molinari, O'Donoghue and Teitelbaum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Unobserved Heterogeneity

Unobserved heterogeneity refers to inter-individual differences that cannot be measured by observables. "The role of unobserved heterogeneity lies at the heart of numerous empirical puzzles and conundrums" (Cameron and Trivedi, 2009). Observed heterogeneity may be adequately accounted for through covariates, but the presence of unobserved heterogeneity implies additional variation that confounds the impact of the observables and may invalidate statistical conclusions. It is important to consider models that incorporate unobserved heterogeneity and allow economic interpretation of what it represents.

Longitudinal, or panel, data have the potential to resolve fundamental issues regarding sources of heterogeneity. Longitudinal data arise when repeated measurements are observed on each cross-sectional unit, thus providing information about individual behavior across individual and across time.[1] Data of this structure offers a way to account for unobserved time-invariant individual-specific effects, i.e. unobserved heterogeneity, through individual-specific effects models. Such models incorporate a time-invariant individual-specific term to capture unobservable effects: as either fixed or random. Random effects models assume the subject-specific effect is an i.i.d. random variable and involves estimating only the parameters of the distribution of the subject-specific term. Fixed effects models involve estimating the subject-specific term as an incidental parameter.

[1]The focus of this dissertation is short panel data, a prevalent form of panel data in microeconomics, where a large cross section of individuals is observed for a short amount of time.

This research accounts for and assesses unobserved heterogeneity through random effects. Random effects can be used as a tool for measuring unobserved heterogeneity as they represent subject-level variation. Random effects capture dependence in longitudinal data while maintaining a parameterization that allows natural economic interpretation. For example, in a model of insurance claim counts, subject-specific random intercepts can be interpreted as the additional level of unobserved "riskiness" associated with the individual. In addition to interpretability, random effects models can easily be extended to multivariate outcomes, a set of jointly dependent outcomes, by imposing correlation between random effects.

1.2 Economic Application: Insurance Data

The motivating data for the research presented in this dissertation was provided by a large U.S. property and casualty insurance company. This data contains information on policy and household characteristics collected annually for multiple lines of coverage. Such characteristics include: claim counts, insurance score (a score derived from information contained in credit reports), driver characteristics, property characteristics, origination dates and cancellation dates. This dataset is unique in that information from each type of coverage can be matched by a unique customer identification number. This matching results in a wealth of data of multivariate structure. In addition to observing policies over time, we observe multivariate dependent variables of interest, specifically a trivariate count outcome (number of home, collision and comprehensive claims) and a bivariate duration outcome (duration of auto and home policies).

The multivariate and longitudinal structure of this data lends itself to models that ad-

dress unobserved heterogeneity, which lead to conclusions about the association between outcomes and unobserved subject-specific effects. We are interested in jointly estimating the association between various outcomes measured over time and policies, specifically claim count and policy duration outcomes. In Chapter 2, we implement a multivariate longitudinal count model that incorporates unobserved heterogeneity to assess the association of inherent time-constant risk characteristics of policyholders as modeled through claim rates for three types of coverage. This gives us insight into the true underlying "riskiness" of the policyholder, specfically into the level of (dis)similarity in how the intrinsic "riskiness" affects claim propensity. In Chapter 3, we detail the advantage that joint estimation of claim counts for the three types of coverage provides in terms of economically significant conclusions. In Chapter 4, we propose a model that jointly assesses the underlying "riskiness" of the policyholder and propensity to maintain a policy in force. This joint model of multivariate longitudinal count and multivariate duration data provides insight into how unobserved policyholder characteristics affect the dropout/cancellation mechanism.

1.3 Joint Modeling

1.3.1 Joint Modeling Approaches

Various approaches for simultaneously analyzing multiple outcomes have been studied in the literature, including: multivariate models, conditional models, dimension reduction, shared parameter models and random effects models. Each of these techniques have advantages and disadvantages, with a common disadvantage of potential compu-

tational complexity when extending to higher dimensional data. This research focuses on models that employ random effects to induce dependence between multiple outcomes. Random effects models allow inference on the original set of outcomes as well as direct marginal inference, consist of separate "univariate" models that are implied by the "multivariate" model, are suitable for different types of outcomes, and impose no dimension restriction. Random effects model can easily be extended to higher dimensions in theory but this advantage does not come without limitations: as the dimension increases, the computational complexity increases. Overall, the random effects approach is very flexible and it is useful and important to develop techniques that overcome computational limitations.

1.3.2 Joint Models for Multivariate Longitudinal Count Data

Methods for accounting for correlation in a single longitudinal count outcome are well-established and straightforward to implement (Cameron and Trivedi, 1998; Winkelmann, 2003). However, often times the researcher is presented with multiple longitudinal outcomes with an underlying relationship that should not be ignored. For example, we are interested in how unobserved individual-specific risk characteristics are related across multiple types of personal insurance coverages. Separate generalized linear mixed models for each of the claim counts can be fit, but a joint model for the multiple claim processes properly addresses our research question. This research focuses on generalized linear mixed models (GLMMs) with correlated random effects that induce marginal association between the multiple claim rates through the joint dependence on the random effects. In such a model, the covariance structure of the random effects tells us about the relationship between the unobserved heterogeneity in the multiple claim count processes.

Maximum likelihood estimation of the multivariate longitudinal GLMM requires distributional specification of the unobserved heterogeneity and is also computationally prohibitive. Specifically, assuming the count outcomes y_{itk} are conditionally Poisson distributed with mean λ_{itk}, it involves maximizing the following marginal likelihood:

$$\prod_{i=1}^{N} \int_{u_{iK}} ... \int_{u_{i1}} \left\{ \prod_{k=1}^{K} \prod_{t=1}^{T_i} e^{-u_{ik}\lambda_{itk}} \frac{(u_{ik}\lambda_{itk})^{y_{itk}}}{y_{itk}!} \right\} g(u_{i1}, ..., u_{iK}) du_{i1} ... du_{iK}$$

where $g(u_{i1}, ..., u_{iK})$ is the multivariate density of the random effects, K is the number of dependent count variables, T_i is the time dimension for subject i, and N is the number of subjects. Pairwise likelihood is one approach to reduce the computational complexity of evaluating and maximizing the above integral (Fieuws and Verbeke, 2006; Fitzmaurice et al., 2009). This method reduces the full likelihood to a composite likelihood that involves fitting all pairwise GLMMs, but this can still be computationally prohibitive and is not robust to misspecification. A semiparametric approach for estimating the association parameters in this joint model is presented in Chapter 2. This robust and computationally feasible method uses the moments implied by the GLMM with correlated random effects in the framework of generalized estimating equations (Liang and Zeger, 1986; Prentice, 1988; Gourieroux et al., 1984a).

Using this semiparametric method, we fit a joint model for multivariate longitudinal insurance claim counts which allows a variety of interesting conclusions regarding intrinsic riskiness of policyholders. Details of these conclusions are presented in Chapter 3. These results are robust to distributional assumptions on the inherently unobservable individual heterogeneity and are not limited by computational complexity due to the large dimension of the insurance data.

1.3.3 Joint Models for Longitudinal Count and Duration Data

The conceptual basis for joint modeling of longitudinal count and duration outcomes is similar to that for multivariate longitudinal count data. Methods for separate analysis of a longitudinal count response, such as GLMMs, and duration outcomes, such as a parametric and Cox proportional hazards model, are well-established and simple to implement. But separate models ignore any correlation between the longitudinal measures and the duration, and thus are inappropriate when the longitudinal outcome is correlated with the time-to-event. For example, in the insurance data, the propensity to cancel an insurance policy may be correlated with the inherent riskiness of the individual: a relationship that should be addressed. We focus on joint models that induce correlation between the longitudinal and duration outcomes using random effects (Wulfsohn and Tsiatis, 1997).

The relation between the duration outcome and the count outcome through shared random effects is a particularly important research question since the method for multivariate longitudinal count data described above assumes no association between the processes, i.e. attrition is random. Joint modeling of these two processes serves many objectives: it characterizes the relationship between the longitudinal process and the duration outcome, accounts for complications of dropout in longitudinal outcomes, and addresses the effect of time-varying covariates in a duration model. Specifically, a joint model of the multivariate longitudinal count process described previously and a multivariate duration outcome - the time to cancellation of home and/or auto policy - can be fit by maximizing the following likelihood:

$$\prod_{i=1}^{N} \int_{u_{iK}} \dots \int_{u_{i1}} \left\{ \prod_{j=1}^{J} f(T_{ij}^{*} | \mathbf{z}_{itj}, u_i, \dots, u_{iK}) \right\} \left\{ \prod_{k=1}^{K} \prod_{t=1}^{T_i} e^{-u_{ik}\lambda_{itk}} \frac{(u_{ik}\lambda_{itk})^{y_{itk}}}{y_{itk}!} \right\} g(u_{i1}, \dots, u_{iK}) du_{i1} \dots du_{iK}$$

where T_{ij}^{*} is the underlying duration for subject i and duration j. Note that this likelihood extends the likelihood for the multivariate longitudinal count model to include an additional term for the duration outcomes. This additional term adds to the computational complexity of the maximum likelihood approach. To overcome this complexity, a method based on approximate variational inference is presented in Chapter 4. This method alleviates the computational problem by making assumptions on the posterior distribution of the random effects to reduce the problem to maximizing a function of closed form.

CHAPTER 2

A SEMIPARAMETRIC APPROACH FOR MULTIVARIATE LONGITUDINAL COUNT DATA

2.1 Introduction

Correlated count data commonly arise in fields such as business, economics and demography through longitudinal studies of a single outcome or cross-sectional studies of multiple outcomes. Methods for accounting for correlation in either of these types of studies are well-established. But researchers may be interested in jointly modeling multiple outcomes measured repeatedly over time. Joint models of multivariate longitudinal data provide a formal framework for answering research questions about the systematic association of the outcomes. This research contributes to the literature on joint modeling of multivariate longitudinal outcomes by providing a semiparametric approach for fitting a correlated random effects model that uses the generalized estimating equation (GEE) framework. The proposed semiparametric method is robust to misspecified distributional assumptions that intrinsically lack verifiability. Our approach also reduces computational complexity, resulting in a substantial computational advantage over comparable likelihood methods. Our method allows for estimation and inference of models that are otherwise computationally prohibitive because of the dimension of the multivariate outcome, the size of the dataset, or the dimension of the covariate or outcome vector.

A standard parametric univariate count model assumes that the count outcome follows a Poisson distribution with an exponential mean function (McCullagh and Nelder, 1989). A simple extension for univariate longitudinal data multiplicatively combines the standard count model with an individual specific term that reflects subject-specific time-

invariant unobserved heterogeneity (Cameron and Trivedi, 1998; Winkelmann, 2003):

$$y_{it}|\mathbf{x}_{it}, u_i \sim Poisson(u_i\lambda_{it}), \qquad i = 1, ..., N,\ t = 1, ..., T_i \tag{2.1}$$

where y_{it} is a scalar count outcome, $\mathbf{x}_{it}$ is a vector of explanatory variables, $\lambda_{it} = exp\left(\mathbf{x}_{it}^T\beta\right)$, and u_i is the individual-specific time-constant term for subject i. Models of this type assume that data are independent over subjects and that correlation over time is adequately controlled for through the subject-specific effects. The introduction of the additional randomness due to the unobserved heterogeneity allows the subject-specific rates to vary in a way that cannot be accounted for by observables.

Estimation of such models may employ either random effects, where the subject-specific effect is assumed to be an i.i.d. random variable and only the parameters of the distribution of the subject specific term are estimated, or fixed effects, where the subject-specific term is estimated as an incidental parameter. This research focuses on random effects models due to flexibility of fitting techniques and weaker assumptions on the form of the association when extended to multiple outcomes, i.e they allow correlation of multiple outcomes through correlated subject-specific effects rather than assuming independence. Moment methods, such as GEE, and likelihood methods, such as generalized linear mixed models (GLMM), can be implemented for these types of models in a way that takes advantage of the information that repeated measures data contains about subject-specific heterogeneity (McCulloch and Searle, 2001). Fitting of this type of model typically involves working with the marginal distribution of $\mathbf{y}_i$ obtained by averaging with respect to the unobserved heterogeneity. In the full likelihood approach, the density of the random effects is specified resulting in an analytical solution only when the conjugate density is used. Intractable densities can be avoided by using moment methods that do not require choosing a distribution for the random effects or the marginal response. Liang and Zeger (1986) and Prentice (1988) propose estimation via GEE based on

marginal moments. The marginal moments are implied by modifying the score equations from the likelihood function produced by the generalized linear model with a weight matrix. The weight matrix is the inverse of the marginal variance matrix which depends on a set of association parameters. If the marginal interpretation of the regression and association parameters is of interest, then the full likelihood approach can be replaced by these moment based methods.

Multivariate cross-section count data arise when a set of jointly dependent outcomes is measured at a fixed point in time. One can work directly with a fully specified multivariate distribution determined from either some decomposition of marginals and conditionals or based on some joint distribution that leads to Poisson marginals (Kocherlakota and Kocherlakota, 1993). Alternatively, as in the univariate longitudinal case, introducing correlated unobserved heterogeneity proves to be a useful and flexible way to induce dependence in a multi-dimensional outcome vector. Full likelihood estimation of multivariate parametric count models, such as the multivariate Poisson-gamma mixture model and the multivariate Poisson-lognormal mixture model, require numerically intensive methods that get harder as the dimension of the outcome vector increases. Gourieroux et al. (1984b) introduced a moment-based procedure for a flexible bivariate count model using both a pairwise shared parameter and a subject-specific unobserved component.

The univariate longitudinal and multivariate cross-sectional models and methods discussed so far concern only a single outcome of interest measured over time or multiple outcomes of interest measured at one point in time for a set of subjects. The underlying concepts of these methods can be extended to multivariate longitudinal data. Specifically, a correlated random effects model is a flexible and useful tool for multivariate longitudinal data. Full likelihood methods of such models are not necessarily feasible and computational problems arise due to complex integration of possibly very high dimen-

sional integrals. Fieuws and Verbeke (2006) propose a pseudo-likelihood method for pairwise fitting of a system of generalized linear mixed models with correlated outcomes and subject specific random effects, a special case of composite marginal likelihood theory (Lindsay, 1988). Even this pairwise model decomposition can be very computationally intensive depending on the distributional assumptions and the dimension of the data. Extending the moment methods commonly used for the estimation of univariate longitudinal and multivariate cross-section count models can overcome the computation burden while maintaining the joint modeling framework.

The methodology proposed in this chapter combines generalized estimating equations and random effects models for multivariate longitudinal data by introducing a specific structure for the weighting matrix used in the estimating equations for the regression parameters that in turn implies a second set of estimating equations for the association parameters. This structure incorporates latent effects to account for any dependence between outcomes within and between time periods through a multivariate relation between the subject-specific random effects. A marginal approach is of particular interest as our research question concerns the association parameters of the subject specific heterogeneity rather than the individual levels.

In this chapter, we show that the semiparametric methodology proposed for estimating the association parameters of the multivariate longitudinal marginal count model is robust to distributional misspecification and computationally feasible with large datasets. Through simulation studies we illustrate the finite sample robustness properties of the semiparametric method and provide evidence that distributional misspecification can have considerable impact on inferential conclusions. Additionally, this moment-based method addresses the computational challenges associated with the correlated random effects count model. In fact, we find our method to run about 25 times faster than com-

parable likelihood-based methods: a computational advantage that makes it feasible to answer our underlying research question about the relation of unobserved heterogeneity in multiple count processes.

The rest of this chapter is organized as follows: Section 2.2 introduces the motivating research question and the insurance data; Section 2.3 describes the model and the semi-parametric method; Section 2.4 presents simulation studies demonstrating finite sample properties; Section 2.5 discusses the main empirical findings; Section 2.6 reviews important empirical considerations; and Section 2.7 concludes.

2.2 Motivating Example: Insurance Data

This research is motivated by an empirical question concerning the association of unobserved heterogeneity in multiple insurance claim count processes. The unobserved heterogeneity in an insurance claim rate model represents the inherent characteristics of the policyholder that affect the claim rate, after accounting for observable characteristics of the policyholder. Unobserved heterogeneity is thus a measure of the unobserved riskiness of the policyholder: a concept that can be assessed through random effects in a count model. We are interested in how unobserved heterogeneity is associated between different types of insurance coverages, i.e. the level of (dis)similarity in how intrinsic riskiness affects claim propensity.

The motivating dataset, acquired from a large U.S. property and casualty insurance company, contains yearly information on the number of claims for multiple lines of personal insurance coverage. This dataset contains household level matched records for home and auto insurance observed over the course of nine years, 1998 – 2006. At the

Table 2.1: Summary Statistics for Insurance Claim Counts

Insurance	Percent Non-Zero						
Type	Home	Collision	Comprehensive	Mean	Variance	Min	Max
Home	7.1	0.8	0.4	.079	.089	0	6
Collision	0.8	9.9	0.5	.107	.111	0	5
Comprehensive	0.4	0.5	3.1	.032	.035	0	5

Note: Includes all 294,917 policy/year observations. Summary statistics vary only slightly by year.

beginning of each claim year, we observe a snapshot of policy and household characteristics, such as insurance score, that are linked to the number of claims filed during the course of the year. The dependent variables of interest are:

$$y_{it1} = \text{number of home claims for policy } i \text{ and time period } t$$
$$y_{it2} = \text{number of collision claims for policy } i \text{ and time period } t$$
$$y_{it3} = \text{number of comprehensive claims for policy } i \text{ and time period } t$$

The unbalanced panel sample of 62,425 policies includes those households that have a complete set of outcomes and covariates for all three coverages at any point in the nine year period, i.e. both home and auto policies in force in any year from 1998 – 2006, for a total of 294,917 observations.[1] About 7%, 10%, 3% of the 294,917 observations for home, collision and comprehensive insurance, respectively, have a positive claim count (see Table 2.1). Under 1% of the observations have a positive claim count for each of the three pairwise combinations of insurance types and only 165 observations have a positive claim count for all three types of insurance in a given year.

[1]The balanced subsample of the insurance data includes the 8,731 policies that have both home and auto policies in force for all nine years, for a total of 78,579 observations.

Separate analysis of this trivariate longitudinal count data is easily pursued due to availability of well-established statistical methodology.[2] While independent modeling of each outcome as a function of relevant covariates provides useful information about the marginal effects of the observables on the claim rate, joint modeling of these trivariate longitudinal count data allows us to address the association structure between different types of claims in the insurance data. We are interested in the association structure since this relation between the random effects, or unobserved heterogeneity, provides insight to the underlying unobserved risk-related characteristics of the policyholder. Pinquet (2012, 1998) describes a similar research question to assess experience rating in French non-life insurance in a multi-equation Poisson model using semiparametric methods based on the same moment-based principles proposed in this work.

2.3 Methodology

2.3.1 Notation and Model

Let y_{itk} denote the k^{th} outcome and $\mathbf{x}_{itk}$ denote the p_kx1 vector of covariates observed for the k^{th} count and the i^{th} subject in time period t, where $i = 1, \ldots, N$, $t = 1, \ldots, T_i$ and $k = 1, \ldots, K$. Let $\mathbf{y}_{ik}$, λ_{ik}, and $\mathbf{x}_{ik}$ denote the T_ix1 vectors and T_ixp_k matrix of all measurements for the k^{th} outcome for the i^{th} subject, e.g. $\mathbf{y}_{ik} = \begin{bmatrix} y_{i1k} & \ldots & y_{iT_ik} \end{bmatrix}^T$. Let $\mathbf{y}_i$, λ_i denote the KT_ix1 vectors of all measurements for the i^{th} subject, e.g. $\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{i1}^T & \ldots & \mathbf{y}_{iK}^T \end{bmatrix}^T$.

To extend the standard Poisson random effects count model from the univariate longi-

[2] Barseghyan et al. (2011) use a univariate Poisson panel regression with gamma distributed random effects to independently model claim rates for home, collision and comprehensive insurance for a sample of this insurance data.

tudinal and multivariate cross-section to the multivariate longitudinal setting, let $\mathbf{u}_i$ be the vector of correlated subject-specific latent effects for subject i with elements $(u_{i1}, \ldots, u_{iK})$. Similar to the univariate longitudinal and multivariate cross-section cases, the model generally assumes:

$$\mathbf{y}_i|\mathbf{x}_i, \mathbf{u}_i = \begin{pmatrix} \mathbf{y}_{i1}|\mathbf{x}_{i1}, u_{i1} \\ \vdots \\ \mathbf{y}_{iK}|\mathbf{x}_{iK}, u_{iK} \end{pmatrix} \sim Poisson \begin{pmatrix} u_{i1}\lambda_{i1} \\ \vdots \\ u_{iK}\lambda_{iK} \end{pmatrix}$$

where $\lambda_{itk} = exp\left(\mathbf{x}_{itk}^T \beta_k\right)$. That is, $\mathbf{y}_i$ follows a Poisson distribution conditional on a set of random effects, a set of covariates and a vector of regression parameters $(\beta_1, ..., \beta_K)$, which includes an intercept, that are common to all subjects.[3] The set of covariates may include an offset, the log length of the risk period, with the associated coefficient constrained to 1. Assume $\mathbf{u}_i$ is a K-dimensional vector with mean one, for identification purposes, and covariance matrix Σ. In this model, the specification of Σ captures all the dependence between repeated outcomes, including the association of outcomes measured at different times. Assuming conditional independence, the marginal density of $\mathbf{y}_i$ can be written as:

$$L_i = \int_{u_{iK}} \ldots \int_{u_{i1}} \left\{ \prod_{k=1}^{K} \prod_{t=1}^{T_i} e^{-u_{ik}\lambda_{itk}} \frac{(u_{ik}\lambda_{itk})^{y_{itk}}}{y_{itk}!} \right\} g(u_{i1}, ..., u_{iK}) du_{i1} ... du_{iK} \tag{2.2}$$

where $g(u_{i1}, ..., u_{iK})$ is the multivariate density of the random effects. An assumption on the joint distribution of the random effects $g(u_{i1}, ..., u_{iK})$ can be imposed. The marginal likelihood L_i involves a possibly high-dimensional integral that may be intractable depending on the specification of the distribution of random effects. Under certain assumptions this likelihood reduces to familiar models that can be easily estimated with maximum likelihood. In the simple univariate longitudinal case, the random effects are independent and

[3] We allow for a different set of regression parameters for each count outcome. This is not necessary, i.e. the constraint $\beta_{k'} = \beta_k$ for $k' \neq k$ may be imposed for shared covariates, but is maintained in this research as we have no reason to impose equivalence of covariate effects for each of the coverage types.

$g(\mathbf{u}_i)$ is typically chosen to be the conjugate gamma distribution or the lognormal distribution. In the case of a Poisson-gamma model, L_i reduces to a product of negative binomial densities. However, for joint modeling of multivariate longitudinal data, one needs to focus on methods that specify a fully multivariate distribution for the random effects.

Distributional assumptions lead to efficiency when they are correctly specified, but the maximum likelihood estimator may be inconsistent if the incorrect distribution is assumed. The latent effects are by definition unobserved, so that any distributional assumption on which the consistency results are based is subjective. This paper provides a method for consistent estimation for all possible distributions of the association parameters of the unobserved effect. In simulation, we show that this robustness property of the semiparametric approach is maintained in finite samples. Also, while composite likelihood methods for this class of generalized linear mixed models can be used to estimate the parameters, computational difficulties arise as the dimension of the random effect vector increases even in seemingly simple cases (Fieuws and Verbeke, 2006). This pairwise likelihood approach sacrifices efficiency for computational gain over full likelihood; in the same vein, the semiparametric approach presented in this paper sacrifices efficiency for a much greater computational gain. In simulation and empirical analysis, we find that the semiparametric method runs about 25 times faster than the pairwise likelihood method. Our moment-based semiparametric approach is both robust to distributional assumptions and significantly reduces the computation burden.

The following formalizes the minimal assumptions associated with this class of multivariate longitudinal count models with multiplicative correlated random effects. These are the assumptions maintained in this research.

Assumption 2.3.1 (Conditional Moments and Model Assumptions)

$$(i)\ E(y_{itk}|\mathbf{x}_{ik}, u_{ik}) = u_{ik}exp(\mathbf{x}_{itk}^T\beta_k) = u_{ik}\lambda_{itk}$$

$$(ii)\ E(y_{itk}|\mathbf{x}_{ik}, u_{ik}) = Var(y_{itk}|\mathbf{x}_{ik}, u_{ik})$$

$$(iii)\ E(u_{ik}|\mathbf{x}_{ik}) = E(u_{ik}) = 1\ and\ Var(u_{i1}, \ldots, u_{iK}|\mathbf{x}_{ik}) = \Sigma$$

$$(iv)\ y_{itk} \perp\!\!\!\perp y_{isl}|(u_{ik}, u_{il})$$

That is, this model assumes a multiplicative subject-specific time-constant random effect for each outcome k, an exponential mean function of the linear predictor $\mathbf{x}_{itk}^T\beta_k$, strictly exogenous covariates, mean independent random effects with a mean of one and KxK covariance matrix Σ and conditional independence. The random effects are assumed to be mean one for identification purposes. By the law of iterated expectations, the first two marginal moments for this class of models can be derived.

Result 2.3.2 (Marginal Moments)

$$(i)\ E(y_{itk}|\mathbf{x}_i) = E(y_{itk}|\mathbf{x}_{itk}) = exp\left(\mathbf{x}_{itk}^T\beta_k\right) = \lambda_{itk}$$

$$(ii)\ V_i \equiv Var(\mathbf{y}_i|\mathbf{x}_i) = diag(\lambda_i^T) + \Sigma \otimes \mathbf{1}_{T_i}\mathbf{1}_{T_i}^T \circ \lambda_i\lambda_i^T$$

where $\circ$ is element-wise multiplication, $\otimes$ is the Kronecker product, and $\mathbf{1}_{T_i}$ is a T_i-dimensional vector of ones.

The semiparametric approach for fitting the multivariate longitudinal count model relies on the moment conditions implied by the marginal mean and variance along with the basic assumptions for multiplicative correlated random effects models.[4] The structure

[4]Assumption 2.3.1*(i)* and Result 2.3.2*(i)* imply that the distinction of estimating a population-averaged model versus a mixed effects model is unimportant. Since only a random intercept is introduced through the log link function, all parameters in β have a marginal interpretation with the exception of the intercept (Aerts et al., 2002). This is evident by the fact that $E(y_{itk}|\mathbf{x}_{itk}) = h(\mathbf{x}_{itk}^T\beta_k)$ and $E(y_{itk}|\mathbf{x}_{itk}, u_{ik}) = h(\mathbf{x}_{itk}^T\beta_k + log(u_{ik}))$ where h is the exponential function.

of the model based variance defined in Result 2.3.2*(ii)* allows the random effects to account for the association between the outcomes within a subject, specifically through the parameters in the off-diagonal blocks of the covariance matrix of the latent effects.

2.3.2 Semiparametric Estimation

The procedure for the semiparametric approach involves iterating between moment-based estimation of the covariance parameters of the random effect, Σ, and moment-based estimation of the regression parameters, β. This approach is an extension of quasi-generalized pseudo maximum likelihood (QGPML) estimators developed by Gourieroux et al. (1984a) and the extended GEE approach developed by Prentice (1988). The QGPML method can be characterized as first order GEE with a specific association structure. Prentice (1988) introduces an extension of first order GEE that utilizes a second set of estimating equations to jointly estimate the association parameters. QGPML can be embedded in the GEE framework resulting in commonly studied consistency and asymptotic results for simultaneous inference on both the regression parameters and the association parameters.

The estimator $(\hat{\beta}, \hat{\Sigma})$ for β and Σ is defined as the solution to:

$$U(\beta, \Sigma) = \sum_{i=1}^{N} \begin{pmatrix} D_i^T & 0 \\ 0 & E_i^T \end{pmatrix} \begin{pmatrix} V_i & 0 \\ 0 & W_i \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_i - \lambda_i \\ \mathbf{R}_i^* - \mathbf{V}_i^* \end{pmatrix} = 0$$

where $D_i = \frac{\partial \lambda_i^T}{\partial \beta} = diag\left[\mathbf{x_{i1}}^T \lambda_{i1}, \ldots, \mathbf{x_{iK}}^T \lambda_{iK}\right]$, V_i is the model based variance matrix as defined in Result 2.3.2*(ii)*, $E_i = \frac{\partial \mathbf{V}_i^{*T}}{\partial \mathbf{\Sigma}^*} = diag\left[\left(\lambda_{i1}\lambda_{i1}^T\right)^*, \ldots, \left(\lambda_{iK}\lambda_{iK}^T\right)^*, \left(\lambda_{i1}\lambda_{i2}^T\right)^*, \ldots, \left(\lambda_{i(K-1)}\lambda_{iK}^T\right)^*\right]$, $W_i = I_{T_iK}$ and R_i is the cross product of residuals. Note that other working association matrices can be substituted into W_i. Define $\mathbf{R}_i^*$, $\mathbf{V}_i^*$ and $\mathbf{\Sigma}^*$ to be the vector of unique elements

of R_i, V_i and Σ, respectively.

The semiparametric estimator for this multivariate longitudinal count model involves a two-step iterative procedure. After finding initial estimates for β via a procedure such as nonlinear least squares (i.e. ignoring any dependence between outcomes), $Var(\mathbf{y}_i)$ can be consistently estimated by the cross product of the residuals, R_i. Define the vector of residuals $\hat{\mathbf{r}}_{ik} = \begin{bmatrix} \hat{r}_{i1k} & \dots & \hat{r}_{iT_ik} \end{bmatrix}^T$ with $\hat{r}_{itk} = y_{itk} - e^{\mathbf{x}_{itk}^T\hat{\beta}_k}$, so that:

$$R_i = \widehat{Var}(\mathbf{y}_i) \quad = \quad \begin{bmatrix} \hat{\mathbf{r}}_{i1} \\ \vdots \\ \hat{\mathbf{r}}_{iK} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{r}}_{i1}^T & \dots & \hat{\mathbf{r}}_{iK}^T \end{bmatrix}$$

Next the empirical variance estimate R_i and the model defined variance structure V_i from Result 2.3.2*(ii)* are used to estimate Σ. Specifically, the relation between R_i and V_i implies $\frac{KT_i(KT_i+1)}{2}$ estimating equations for Σ from the distinct elements of the two matrices. For example, the $(1,1)$ element implies a relation between $\hat{r}_{i11}^2$ and $\hat{\lambda}_{i11} + \sigma_1^2\hat{\lambda}_{i11}^2$. These estimating equations define the estimator $\hat{\mathbf{\Sigma}}^*$ for $\mathbf{\Sigma}^*$:

$$U(\mathbf{\Sigma}^*) = \sum_{i=1}^{N} E_i^T(\hat{\beta}) W_i^{-1} \left(\mathbf{R}_i^*(\hat{\beta}) - \mathbf{V}_i^*(\hat{\beta}, \mathbf{\Sigma}^*) \right) = 0$$

The diagonal structure of W_i implies the working assumption that the higher order associations are equal to zero. An estimate for $\mathbf{V_i^*}$ can now be obtained by plugging $\hat{\mathbf{\Sigma}}^*$ into the model defined variance structure from Result 2.3.2*(ii)*. The estimator $\hat{\beta}$ for β is then found as the solution to:

$$U(\beta) = \sum_{i=1}^{N} D_i^T(\beta) \hat{V}_i^{-1}(\hat{\beta}, \hat{\mathbf{\Sigma}}^*) \left(\mathbf{y}_i - \lambda_i(\beta) \right) = 0$$

The roots of the estimating equations $U(\beta, \Sigma)$ are solved for via an iterative procedure, updated at each iteration by the previous value of the $\sqrt{N}$-consistent estimator of β given Σ and the $\sqrt{N}$-consistent estimator of Σ given β, until convergence. Note that each set of estimating equations can be solved via the Newton-Raphson method.

Result 2.3.3 (Consistency of $(\hat{\beta}, \hat{\Sigma}^*)$) *Given Result 2.3.2(i) and under regularity conditions outlined in the Appendix, the estimator that solves $U(\beta; \hat{\Sigma}^*) = 0$ is a consistent estimator for the true parameter β. Additionally, given Result 2.3.2(ii), the estimator that solves $U(\mathbf{\Sigma}^*; \hat{\beta}) = 0$ is a consistent estimator for the true parameter $\mathbf{\Sigma}^*$.*

These consistency results follow from the work of Liang and Zeger (1986), Gourieroux et al. (1984b) and properties of two-step M-estimators outlined in Wooldridge (2001). See the Appendix for more details. Generally, consistent and asymptotically normal estimators of β and Σ are obtained from these sets of equations as long as the first and second order marginal moment models are correctly specified. While the sets of estimating equations for the regression parameters, β, and association parameters, Σ, are not assumed to be independent, this procedure operates as if β and Σ are orthogonal to one another. This procedural property transforms the joint estimation procedure into a two step procedure which implies a consistent $\hat{\beta}$ even if the variance structure is misspecified. This can lead to a loss in efficiency.

2.3.3 Semiparametric Inference

Asymptotic results for the joint distribution of the semiparametric estimators follow from general properties of two-step M-estimation described in Wooldridge (2001). Prentice (1988) developed the joint asymptotic distribution of $\sqrt{N}(\beta - \hat{\beta})$ and $\sqrt{N}(\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}^*)$ specific to the GEE framework. The result that follows accounts for the adjustment necessary in a two-stage estimator.

Result 2.3.4 (Joint Asymptotic Variance) *Given the functional forms specified in Result 2.3.2 and under regularity conditions outlined in the Appendix, the joint asymptotic distribution of* $\sqrt{N}(\beta - \hat{\beta})$ *and* $\sqrt{N}(\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}^*)$ *is multivariate normal with mean zero and covariance matrix*

$$N\begin{pmatrix} A & 0 \\ B & C \end{pmatrix}\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}\begin{pmatrix} A & B^T \\ 0 & C \end{pmatrix}$$

where

$$A = \left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1}$$
$$B = \left(\sum_{i=1}^{N} E_i^T W_i^{-1} E_i\right)^{-1}\left(\sum_{i=1}^{N} E_i^T W_i^{-1} \frac{\partial(\mathbf{R}_i^* - \mathbf{V}_i^*)}{\partial\beta}\right)^{-1}\left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1}$$
$$C = \left(\sum_{i=1}^{N} E_i^T W_i^{-1} E_i\right)^{-1}$$

The Ω *matrix is defined by*

$$\Omega_{11} = \left(\sum_{i=1}^{N} D_i^T V_i^{-1} cov(\mathbf{y}_i) V_i^{-1} D_i\right)^{-1}$$
$$\Omega_{12} = \left(\sum_{i=1}^{N} D_i^T V_i^{-1} cov(\mathbf{y}_i, (\mathbf{R}_i^* - \mathbf{V}_i^*)) W_i^{-1} E_i\right)^{-1}$$
$$\Omega_{22} = \left(\sum_{i=1}^{N} E_i^T W_i^{-1} cov((\mathbf{R}_i^* - \mathbf{V}_i^*)) W_i^{-1} E_i\right)^{-1}.$$

These quantities can be consistently estimated by evaluating at the final parameter estimates for β and Σ and replacing:

$$cov(\mathbf{y}_i) \text{ with } \left(\mathbf{y}_i - \hat{\lambda}_i\right)\left(\mathbf{y}_i - \hat{\lambda}_i\right)^T$$
$$cov(\mathbf{R}_i^* - \mathbf{V}^*{}_i) \text{ with } \left(\mathbf{R}_i^* - \hat{\mathbf{V}}_i^*\right)\left(\mathbf{R}_i^* - \hat{\mathbf{V}}_i^*\right)^T$$
$$cov\left(\mathbf{y}_i, (\mathbf{R}_i^* - \mathbf{V}_i^*)\right) \text{ with } \left(\mathbf{y}_i - \hat{\lambda}_i\right)\left(\mathbf{R}_i^* - \hat{\mathbf{V}}_i^*\right)^T$$

The empirically corrected sandwich estimator of the variance for $\hat{\beta}$ reduces to the model based estimator, $N\left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1}$, in the case that the variance structure V_i in Result 2.3.2*(ii)* is correctly specified. A simpler form of the joint asymptotic variance matrix might assume $B = 0$, that is the covariance between the two sets of estimating equations is zero. With this simplification any dependence between the set of estimating equations for the marginal responses and the set of estimating equations for the association is ignored in the asymptotic variance matrix. This simplification is not assumed in this research, though in simulations and data analysis this covariance adjustment has very little effect.

2.3.4 Alternative Likelihood Estimation and Inference

Given the computational complexities of a full likelihood approach associated with evaluating L_i, Fieuws and Verbeke (2006) propose reducing the dimensionality of joint generalized mixed model by using a pairwise likelihood approach that involves fitting all bivariate pairs of mixed models and combining estimates using pseudo-likelihood theory. This corresponds to the following pseudo likelihood:

$$\prod_{l=1}^{K-1} \prod_{m=l+1}^{K} \left(\prod_{i=1}^{N} L_{i,lm}\left(\mathbf{y}_{il}, \mathbf{y}_{im} | \Theta_{lm}\right) \right)$$

In the context of the multivariate longitudinal count model, each $L_{i,lm}\left(y_{il}, y_{im} | \Theta_{lm}\right)$ corresponds to the two-dimensional version of L_i. Each of these pairwise estimates is consistent and asymptotically normal. Non-unique parameter estimates are averaged together, maintaining the same optimal asymptotic properties. Standard errors are obtained via a sandwich estimator of variance.

2.4 Simulation Studies

This section presents Monte Carlo simulation studies of the finite sample properties of the semiparametric estimator. We assess the bias, variability and root mean squared error (RMSE) of the semiparametric approach compared to the pairwise likelihood approach under different distributional assumptions of the random effects.

2.4.1 Simulation Design

Set the number of outcomes $K = 3$, the number of subjects $N = 3600$, the maximum number of time periods $\max(T_i) = 9$, each $\mathbf{x}_{itk} = [1, x_{itk}]$,[5] β_k the set consisting of an intercept parameter β_{k0} and a slope parameter β_{k1}, and $\mathbf{u}_i$ the 3x1 vector of random effects distributed according to some distribution g with mean 1 and covariance matrix Σ. An offset, the log length of the risk period, is included to account for the unbalanced structure with non-informative dropout and random censoring. The following data generating process is considered:

$$y_{itk} \sim Poisson(\lambda_{itk})$$

$$\lambda_{itk} = u_{ik} e^{\beta_{k0} + x_{itk}\beta_{k1} + \text{offset}_{itk}}$$

$$\mathbf{u}_i = (u_{i1}, u_{i2}, u_{i3}) \sim g(1, \Sigma)$$

In the base case simulation study the model distribution assumptions for the likelihood approach are correctly specified, i.e. the true random effect distribution is multivariate lognormal. Note that even in this case the pairwise likelihood method is slightly

[5] In general, and otherwise in the paper, the first element of $\mathbf{x}_{itk}$ is assumed to be 1 to account for the intercept.

misspecified as it assumes pairs of bivariate lognormal distributions. This base case is compared to simulation studies conducted to assess the semiparametric and pairwise likelihood methods when the random effect distributional assumptions are misspecified, i.e. non-normal. In particular, we define the random effect distribution g to be either a multivariate gamma distribution via a Gaussian copula or a mixture of multivariate lognormal distributions. Specifically, the following latent effect distributions are considered:

$$\mathbf{u}_i \sim \text{lnN}_K(1, \Sigma)$$

$$\mathbf{u}_i \sim C\left(\Gamma(\sigma_{11}^{-1}, \sigma_{11}), \ldots, \Gamma(\sigma_{KK}^{-1}, \sigma_{KK}); \rho(\Sigma)\right)$$

$$\mathbf{u}_i \sim \begin{cases} \text{lnN}_K(\mu_1, \Sigma_1) \text{ with probability } .5 \\ \text{lnN}_K(\mu_2, \Sigma_2) \text{ with probability } .5 \end{cases}$$

where C is the Gaussian copula, $\mu_1 = 1 - \delta$, $\mu_2 = 1 + \delta$, $\delta = [.4, .3, .3]$, $\Sigma_1 = 1.9\Sigma$ and $\Sigma_2 = .1\Sigma$. The multivariate gamma and mixture of multivariate lognormal distributions are used to capture different departures from the multivariate lognormal distribution: skewness and bimodality. The degree of the departure depends on the given variance of the random effects. The two-dimensional kernel densities of the three random effects associated with the three simulated outcomes for one replication are plotted in Figure 2.1.

The covariate x_{itk} is generated as draws from the empirical distribution of the insurance score variable in the insurance data sample and the simulation study slope parameters are set based on estimates observed in a model of the insurance claim data that conditions on a reduced set of covariates. Two sets of intercept parameters are defined to assess effects of different levels of the sparsity in the count outcome. Simulation scenario H, the high mean case, corresponds to marginal means of about 1.4, 1.8 and .6 for the three count outcomes. Simulation scenario L, the low mean case, corresponds to marginal

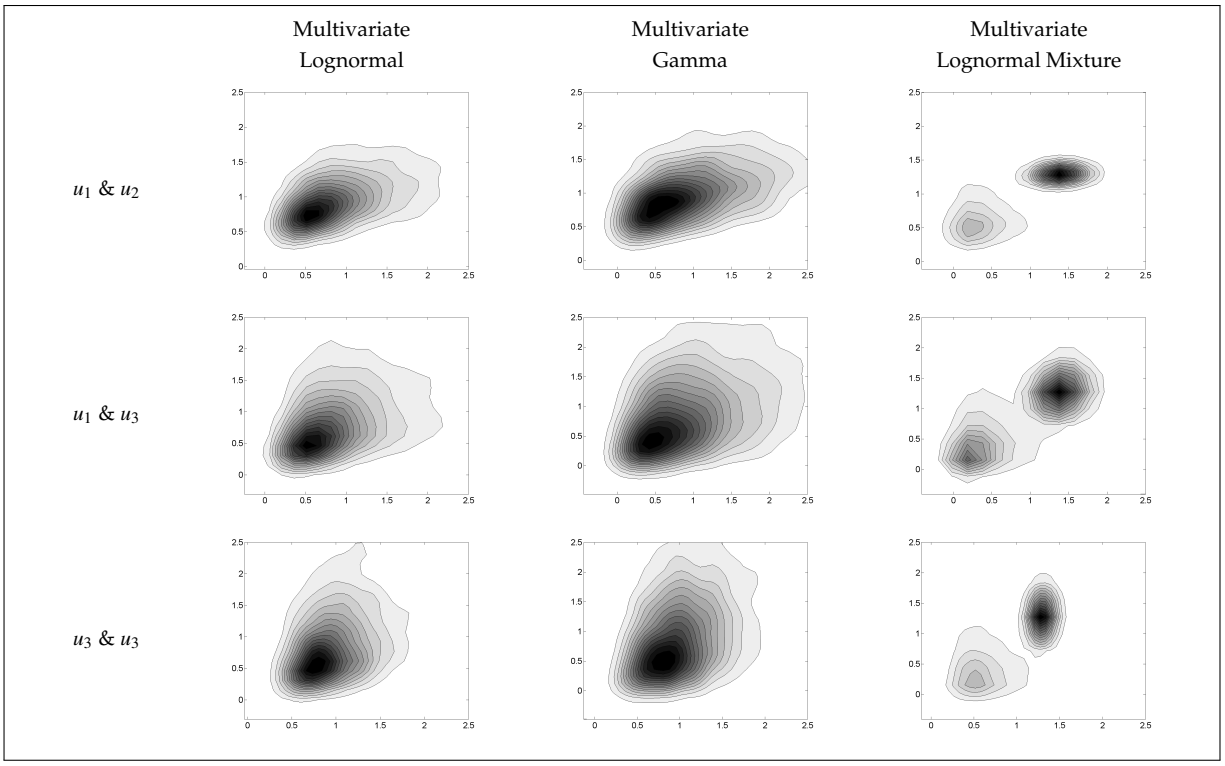

Figure 2.1: Two-Dimensional Kernel Density Plots of Simulated Random Effects

means of about .08, .11 and .03 for the three count outcomes.

Scenario H:

$$\beta_{10}^{H} = \ 1.56,\ \beta_{20}^{H} = \ 1.36,\ \beta_{30}^{H} = \ .063$$

$$\beta_{11} = -.016,\ \beta_{21} = -.010,\ \beta_{31} = -.008$$

Scenario L:

$$\beta_{10}^{L} = -1.25,\ \beta_{20}^{L} = -1.45,\ \beta_{30}^{L} = -2.75$$

$$\beta_{11} = -.016,\ \beta_{21} = -.010,\ \beta_{31} = -.008$$

In order to maintain a comparable level of mean square error to that of the insurance data in simulation, the small simulated sample size is offset by a larger marginal mean

in scenario H. This adjustment roughly corresponds to fixing the number of non-zero counts: with sparse counts in scenario L and non-sparse counts in scenario H. Since simulation scenario H reflects the level of claim "information" in the insurance data through a comparable estimated level of variability, we will focus on this scenario. In contrast, simulation scenario L, reflects the marginal means observed in the insurance data presented in Table 2.1, but with a drastically reduced subject size N in the simulation study: $3,600$ vs $62,435$. The relationship between the number of observational units N, the number of time periods T_i and the marginal mean has significant impact in assessing the finite sample properties of the estimator. This complication is discussed in more detail in Section 6.1. The variance and covariance parameters in the simulation study are set based on estimates observed in a model of the insurance claim data that conditions on a reduced set of covariates:

$$\Sigma = \begin{bmatrix} .469 & .139 & .221 \\ .139 & .155 & .105 \\ .221 & .105 & .556 \end{bmatrix}$$

2.4.2 Simulation Results

The simulation studies show that the pairwise likelihood estimator of the covariance matrix of the unobserved heterogeneity is strongly biased upward when the random effect distribution is misspecified and the count outcome is not sparse (see Table 2.2). The empirical bias associated with the variance parameters is tens to thousands times larger and the bias for the covariance parameters is about one to hundreds times larger using the pairwise likelihood method as compared to the semiparametric method, with the bias for the pairwise likelihood estimator ranging from 6 to 90% of the true value of the parameter. In the correctly specified case the pairwise likelihood and semiparametric estimators

perform similarly well in terms of level of bias. The bias of the pairwise likelihood estimator is less pronounced in the low mean case where the count outcome is sparse. In this scenario, regardless of the distributional misspecification, both the pairwise likelihood and semiparametric estimators of Σ result in a bias of no more than about 6% of the true value of the parameter. In terms of the bias of the variance and covariance parameters, the semiparametric approach is more robust to departures from lognormality when enough "information" is present in the data and similarly robust to distributional misspecification as the pairwise likelihood with sparse outcome data. Regardless of the simulation scenario, the bias of the regression parameters β are found to be similarly robust for both methods with small overall levels of bias.[6]

To evaluate efficiency loss, the empirical relative efficiency (RE) in Table 2.2 is calculated as the ratio of the empirical MSE of the pairwise likelihood estimator and the empirical MSE of the semiparametric estimator. The semiparametric estimator of the covariance matrix of the unobserved heterogeneity is generally much more efficient than the pairwise likelihood estimator when the random effects distribution is misspecified, with the RE ranging from 2 to 44 for the variance parameters and 1 to 125 for the covariance parameters in the high mean setting, with a few exceptions. The loss of efficiency for the pairwise likelihood estimator as compared to the semiparametric estimator under misspecification is not apparent when the sparse count setting is simulated. In this case, generally the RE ranges from about .7 to 1 regardless of misspecification indicating that while both methods are robust to misspecification in this setting, the semiparametric approach exhibits only a small loss in efficiency, with a few exceptions. Figure 2.2 graph-

[6]Note that in order to compare the methods, the estimates and standard errors of the random effect covariance parameters from the pairwise likelihood approach must be transformed since the random effects are parameterized to enter additively in the linear predictor. The transform of the estimates simply involves the relation between the multivariate normal and multivariate lognormal. The standard errors for the pairwise likelihood approach are transformed using the Delta method. The intercepts have been adjusted for the mean shift so that the results reflect the assumption that $E(u_{ik}) = 1$.

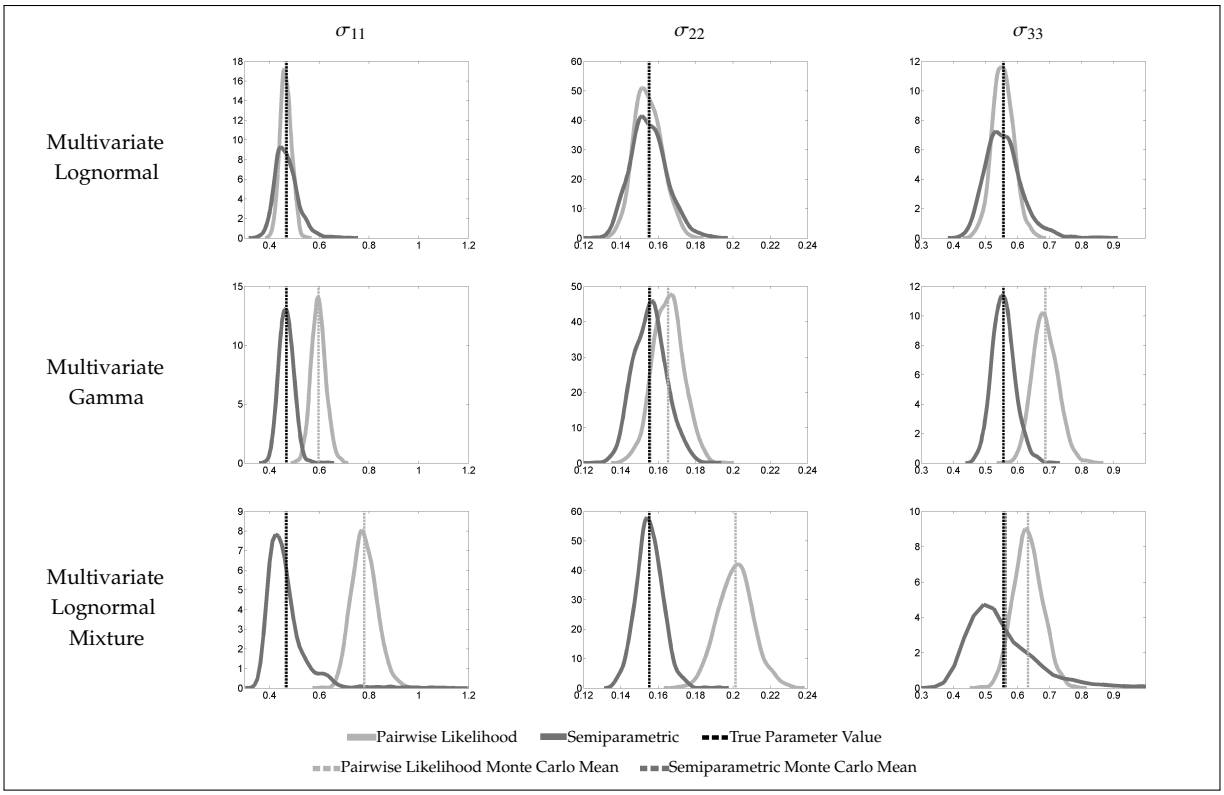

Figure 2.2: Kernel Density Plots of Variance Parameter Estimates from High Mean Simulation Study

ically depicts how both bias and variability effect the empirical efficiency of the estimator for the variance parameters in the non-sparse count case.

The empirical relative efficiency for the regression parameters β are found to be relatively stable around 1.0 in all simulated scenarios. Note that the simulation studies only investigate properties with respect to changes in the random effect distributions, not changes in the marginal response distribution. In both methods, the marginal mean specification requires the mean-variance relationship of the Poisson distribution. If this relation is violated then the semiparametric and pairwise likelihood methods are misspecified.

Table 2.2: Estimation Results from Simulation Study:
Semiparametric and Pairwise Likelihood Method

$K = 3$, $N = 3600$, $\max(T_i) = 9$, scalar x_{itk} plus an intercept, 1000 replications

	Scenario L: Low Mean								Scenario H: High Mean							
		Semiparametric			Pairwise Likelihood					Semiparametric			Pairwise Likelihood			
	θ^L	$\hat{\theta}_{MC}$	Bias	RMSE	$\hat{\theta}_{MC}$	Bias	RMSE	RE^L	θ^H	$\hat{\theta}_{MC}$	Bias	RMSE	$\hat{\theta}_{MC}$	Bias	RMSE	RE^H
Multivariate Lognormal																
σ_{11}	0.469	0.467	-0.002	0.101	0.467	-0.002	0.083	0.674	0.469	0.470	0.001	0.047	0.466	-0.003	0.023	0.239
σ_{22}	0.155	0.155	-0.001	0.052	0.156	0.000	0.049	0.860	0.155	0.155	0.000	0.010	0.155	-0.001	0.007	0.601
σ_{33}	0.556	0.554	-0.002	0.202	0.554	-0.003	0.180	0.794	0.556	0.557	0.001	0.058	0.555	-0.001	0.033	0.329
σ_{12}	0.139	0.140	0.001	0.042	0.140	0.001	0.040	0.908	0.139	0.139	0.000	0.013	0.138	-0.001	0.008	0.439
σ_{13}	0.221	0.219	-0.002	0.084	0.219	-0.003	0.078	0.844	0.221	0.222	0.001	0.028	0.220	-0.001	0.016	0.302
σ_{23}	0.105	0.103	-0.003	0.059	0.102	-0.003	0.057	0.940	0.105	0.105	0.000	0.012	0.105	0.000	0.009	0.597
β_{10}	-1.250	-1.261	-0.011	0.219	-1.261	-0.011	0.219	0.997	1.560	1.561	-0.002	0.107	1.561	-0.002	0.109	1.037
β_{11}	-0.016	-0.016	0.000	0.003	-0.016	0.000	0.003	0.998	-0.016	-0.016	0.000	0.001	-0.016	0.000	0.001	1.025
β_{20}	-1.450	-1.451	-0.001	0.190	-1.451	-0.001	0.190	1.000	1.360	1.365	0.001	0.068	1.364	0.001	0.068	1.001
β_{21}	-0.010	-0.010	0.000	0.002	-0.010	0.000	0.002	1.000	-0.010	-0.010	0.000	0.001	-0.010	0.000	0.001	1.000
β_{30}	-2.750	-2.744	0.006	0.339	-2.743	0.007	0.339	1.002	0.063	0.059	-0.005	0.130	0.059	-0.005	0.127	0.957
β_{31}	-0.008	-0.008	0.000	0.004	-0.008	0.000	0.004	1.003	-0.008	-0.008	0.000	0.002	-0.008	0.000	0.002	0.955
Multivariate Gamma																
σ_{11}	0.469	0.467	-0.002	0.087	0.484	0.016	0.081	0.851	0.469	0.468	-0.001	0.029	0.597	0.129	0.132	20.097
σ_{22}	0.155	0.155	0.000	0.051	0.157	0.002	0.048	0.860	0.155	0.156	0.000	0.009	0.165	0.010	0.013	2.173
σ_{33}	0.556	0.549	-0.007	0.194	0.560	0.004	0.186	0.916	0.556	0.556	0.000	0.035	0.687	0.131	0.136	14.949
σ_{12}	0.139	0.134	-0.005	0.042	0.136	-0.003	0.039	0.896	0.139	0.134	-0.005	0.012	0.152	0.013	0.016	1.884
σ_{13}	0.221	0.208	-0.014	0.077	0.215	-0.007	0.073	0.907	0.221	0.210	-0.012	0.024	0.253	0.032	0.037	2.415
σ_{23}	0.105	0.100	-0.005	0.058	0.101	-0.004	0.057	0.955	0.105	0.100	-0.005	0.012	0.111	0.006	0.012	0.932
β_{10}	-1.250	-1.249	0.001	0.217	-1.248	0.002	0.217	1.006	1.563	1.569	0.006	0.104	1.577	0.013	0.114	1.207
β_{11}	-0.016	-0.016	0.000	0.003	-0.016	0.000	0.003	1.007	-0.016	-0.016	0.000	0.001	-0.016	0.000	0.001	1.173
β_{20}	-1.450	-1.446	0.004	0.183	-1.446	0.004	0.183	1.000	1.363	1.364	0.001	0.068	1.364	0.001	0.069	1.018
β_{21}	-0.010	-0.010	0.000	0.002	-0.010	0.000	0.002	1.000	-0.010	-0.010	0.000	0.001	-0.010	0.000	0.001	1.020

Continued on next page...

	Scenario L: Low Mean								Scenario H: High Mean							
		Semiparametric			Pairwise Likelihood					Semiparametric			Pairwise Likelihood			
	θ^L	$\hat{\theta}_{MC}$	Bias	RMSE	$\hat{\theta}_{MC}$	Bias	RMSE	RE^L	θ^H	$\hat{\theta}_{MC}$	Bias	RMSE	$\hat{\theta}_{MC}$	Bias	RMSE	RE^H
β_{30}	-2.750	-2.747	0.003	0.329	-2.746	0.004	0.330	1.002	0.063	0.064	0.000	0.129	0.070	0.007	0.133	1.062
β_{31}	-0.008	-0.008	0.000	0.004	-0.008	0.000	0.004	1.003	-0.008	-0.008	0.000	0.002	-0.008	0.000	0.002	1.058
Mixture of Multivariate Lognormal																
σ_{11}	0.469	0.466	-0.003	0.138	0.468	-0.001	0.091	0.433	0.469	0.466	-0.003	0.087	0.782	0.313	0.317	13.280
σ_{22}	0.155	0.153	-0.002	0.049	0.155	0.000	0.045	0.857	0.155	0.155	0.000	0.007	0.201	0.046	0.047	43.791
σ_{33}	0.556	0.558	0.002	0.317	0.541	-0.015	0.202	0.407	0.556	0.563	0.007	0.169	0.633	0.077	0.088	0.273
σ_{12}	0.139	0.139	0.000	0.040	0.147	0.008	0.040	1.010	0.139	0.138	0.000	0.009	0.243	0.104	0.105	124.252
σ_{13}	0.221	0.221	0.000	0.078	0.231	0.009	0.074	0.904	0.221	0.222	0.001	0.029	0.418	0.197	0.199	45.679
σ_{23}	0.105	0.105	0.000	0.056	0.109	0.004	0.055	0.958	0.105	0.105	0.000	0.011	0.168	0.063	0.064	36.106
β_{10}	-1.250	-1.251	-0.001	0.215	-1.250	0.000	0.215	0.999	1.563	1.563	0.000	0.104	1.569	0.005	0.116	1.249
β_{11}	-0.016	-0.016	0.000	0.003	-0.016	0.000	0.003	0.998	-0.016	-0.016	0.000	0.001	-0.016	0.000	0.002	1.296
β_{20}	-1.450	-1.453	-0.003	0.187	-1.453	-0.003	0.188	1.003	1.363	1.362	-0.001	0.069	1.362	-0.001	0.073	1.121
β_{21}	-0.010	-0.010	0.000	0.002	-0.010	0.000	0.002	1.003	-0.010	-0.010	0.000	0.001	-0.010	0.000	0.001	1.129
β_{30}	-2.750	-2.746	0.004	0.324	-2.745	0.005	0.324	0.998	0.063	0.063	0.000	0.128	0.068	0.005	0.126	0.968
β_{31}	-0.008	-0.008	0.000	0.004	-0.008	0.000	0.004	0.999	-0.008	-0.008	0.000	0.002	-0.008	0.000	0.002	0.962

Note: $\hat{\theta}_{MC} = .001 \sum_{r=1}^{1000} \hat{\theta}^{(r)}$ is the Monte Carlo estimate of the parameter, bias is the difference between the true value θ and $\hat{\theta}_{MC}$, $RMSE = \sqrt{.001 \sum_{r=1}^{1000} \left(\hat{\theta}^{(r)} - \theta\right)^2}$ is the root mean squared error, and $RE = \frac{MSE_{pair}}{MSE_{semi}}$ is the relative efficiency.

The estimate of the asymptotic variance of $\hat{\beta}$ and $\hat{\Sigma}$ generally accurately reflects the sampling variability in both the pairwise likelihood and semiparametric estimates. Table 2.3 indicates this, as the ratio of the mean of the estimated standard error, $\overline{\hat{se}(\hat{\theta})}$, to the Monte Carlo estimate of the standard error, $se(\hat{\theta})$, is close to 1 in all simulation settings, except for the semiparametric estimator of the standard error associated with the variance parameters in the low mean case. In this case, the sandwich estimator of the asymptotic variance of $\hat{\Sigma}$ overestimates the sampling variability, which is also reflected in the coverage probabilities. Generally, these estimates of the sampling variability of $\hat{\beta}$ and $\hat{\Sigma}$ lead to coverage probabilities of close to .95 for a 95% nominal confidence interval. However, in the non-sparse misspecified simulation scenarios, the large bias of the pairwise likelihood estimator of Σ drives the coverage probabilities to be as low as .002 for the variance parameters and 0 for the covariance parameters. In these cases, the bias of the pairwise likelihood estimator for Σ is so severe that the advantage of precision of the pairwise likelihood estimator is irrelevant.

The pairwise likelihood approach results in a potentially seriously biased estimator that, while more precise than the semiparametric estimator in the correctly specified case, can lead to incorrect conclusions about the covariance matrix of the unobserved heterogeneity in the multivariate longitudinal count model. These simulation studies indicate that under the chosen simulated settings, which reflect the properties of the insurance data, the semiparametric approach is robust to distributional misspecification. This robustness property makes the semiparametric approach more desirable than the pairwise likelihood approach, particularly when the count outcome is not sparse.

Table 2.3: Variability Results from Simulation Study:
Semiparametric and Pairwise Likelihood Method

$K = 3$, $N = 3600$, $\max(T_i) = 9$, scalar x_{itk} plus an intercept, 1000 replications

	Scenario L: Low Mean								Scenario H: High Mean							
	Semiparametric				Pairwise Likelihood				Semiparametric				Pairwise Likelihood			
Parameter	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP
Multivariate Lognormal																
σ_{11}	0.101	0.159	1.576	0.998	0.083	0.083	1.001	0.945	0.047	0.048	1.013	0.943	0.023	0.023	0.988	0.923
σ_{22}	0.052	0.084	1.611	0.993	0.049	0.047	0.969	0.929	0.010	0.011	1.144	0.970	0.007	0.007	1.005	0.952
σ_{33}	0.202	0.436	2.159	1.000	0.180	0.183	1.018	0.947	0.058	0.063	1.085	0.959	0.033	0.033	0.995	0.939
σ_{12}	0.042	0.044	1.042	0.964	0.040	0.040	0.995	0.944	0.013	0.013	1.042	0.949	0.008	0.008	1.005	0.939
σ_{13}	0.084	0.082	0.974	0.942	0.078	0.076	0.973	0.946	0.028	0.027	0.948	0.943	0.015	0.016	1.033	0.950
σ_{23}	0.059	0.060	1.019	0.947	0.057	0.057	1.004	0.937	0.012	0.013	1.037	0.950	0.009	0.009	0.991	0.941
β_{10}	0.219	0.223	1.018	0.956	0.219	0.223	1.019	0.947	0.107	0.106	0.996	0.945	0.109	0.106	0.974	0.941
β_{11}	0.003	0.003	1.020	0.956	0.003	0.003	1.015	0.948	0.001	0.001	0.997	0.943	0.001	0.001	0.980	0.939
β_{20}	0.190	0.186	0.976	0.941	0.190	0.186	0.980	0.939	0.068	0.070	1.028	0.957	0.068	0.070	1.021	0.955
β_{21}	0.002	0.002	0.979	0.939	0.002	0.002	0.976	0.936	0.001	0.001	1.035	0.955	0.001	0.001	1.029	0.955
β_{30}	0.339	0.332	0.979	0.942	0.339	0.336	0.991	0.942	0.130	0.130	0.999	0.940	0.127	0.127	0.993	0.948
β_{31}	0.004	0.004	0.978	0.941	0.004	0.004	0.972	0.939	0.002	0.002	0.996	0.939	0.002	0.002	0.991	0.943
Multivariate Gamma																
σ_{11}	0.087	0.149	1.711	0.997	0.079	0.083	1.048	0.955	0.029	0.033	1.113	0.956	0.029	0.028	0.976	0.003
σ_{22}	0.051	0.083	1.624	0.999	0.048	0.051	1.061	0.937	0.009	0.009	1.091	0.969	0.008	0.008	0.977	0.774
σ_{33}	0.194	0.425	2.186	1.000	0.186	0.184	0.987	0.934	0.035	0.044	1.242	0.980	0.039	0.039	1.014	0.052
σ_{12}	0.041	0.042	1.027	0.950	0.039	0.040	1.013	0.943	0.011	0.011	1.025	0.914	0.010	0.009	0.974	0.729
σ_{13}	0.075	0.078	1.034	0.944	0.073	0.108	1.492	0.951	0.021	0.021	0.997	0.882	0.019	0.019	0.978	0.627
σ_{23}	0.058	0.059	1.020	0.952	0.057	0.150	2.646	0.937	0.011	0.011	1.026	0.924	0.010	0.010	1.010	0.912
β_{10}	0.217	0.222	1.026	0.960	0.217	0.224	1.030	0.962	0.104	0.106	1.023	0.960	0.113	0.116	1.020	0.940
β_{11}	0.003	0.003	1.029	0.963	0.003	0.003	1.023	0.965	0.001	0.001	1.029	0.960	0.001	0.001	1.030	0.940
β_{20}	0.183	0.186	1.012	0.957	0.183	0.190	1.038	0.956	0.068	0.071	1.030	0.965	0.069	0.071	1.034	0.956

Continued on next page...

	Scenario L: Low Mean								Scenario H: High Mean							
	Semiparametric				Pairwise Likelihood				Semiparametric				Pairwise Likelihood			
Parameter	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	$\frac{\overline{\hat{se}(\hat{\theta})}}{se(\hat{\theta})}$	CP
β_{21}	0.002	0.002	1.010	0.959	0.002	0.002	1.019	0.958	0.001	0.001	1.026	0.959	0.001	0.001	1.028	0.956
β_{30}	0.329	0.332	1.009	0.947	0.330	0.342	1.038	0.950	0.129	0.130	1.014	0.957	0.132	0.134	1.013	0.953
β_{31}	0.004	0.004	1.011	0.953	0.004	0.004	1.019	0.950	0.002	0.002	1.009	0.948	0.002	0.002	1.007	0.955
Mixture of Multivariate Lognormal																
σ_{11}	0.138	0.168	1.214	0.992	0.091	0.086	0.945	0.923	0.087	0.062	0.713	0.793	0.050	0.053	1.078	0.002
σ_{22}	0.049	0.082	1.664	0.999	0.045	0.046	1.002	0.944	0.007	0.007	1.012	0.955	0.009	0.010	1.091	0.003
σ_{33}	0.317	0.460	1.450	0.999	0.202	0.189	0.935	0.910	0.169	0.110	0.650	0.838	0.044	0.046	1.036	0.631
σ_{12}	0.040	0.040	0.999	0.946	0.039	0.039	0.982	0.940	0.009	0.009	0.913	0.936	0.014	0.017	1.210	0.002
σ_{13}	0.078	0.078	1.005	0.958	0.073	0.073	1.000	0.941	0.029	0.025	0.839	0.919	0.029	0.029	1.000	0.003
σ_{23}	0.056	0.057	1.020	0.960	0.055	0.056	1.021	0.950	0.011	0.010	0.930	0.934	0.012	0.012	0.994	0.000
β_{10}	0.215	0.223	1.033	0.951	0.215	0.223	1.034	0.952	0.104	0.105	1.018	0.953	0.116	0.125	1.078	0.959
β_{11}	0.003	0.003	1.036	0.955	0.003	0.003	1.029	0.953	0.001	0.001	1.016	0.952	0.001	0.002	1.066	0.945
β_{20}	0.187	0.186	0.991	0.949	0.188	0.186	0.992	0.950	0.069	0.071	1.023	0.951	0.073	0.075	1.033	0.953
β_{21}	0.002	0.002	0.990	0.947	0.002	0.002	0.985	0.946	0.001	0.001	1.016	0.951	0.001	0.001	1.022	0.947
β_{30}	0.324	0.333	1.026	0.952	0.324	0.335	1.035	0.952	0.128	0.130	1.016	0.943	0.126	0.131	1.040	0.958
β_{31}	0.004	0.004	1.024	0.953	0.004	0.004	1.014	0.950	0.002	0.002	1.010	0.945	0.002	0.002	1.032	0.955

Note: $se(\hat{\theta})$ is the Monte Carlo standard deviation of $\hat{\theta}$, $\overline{\hat{se}(\hat{\theta})}$ is the mean of the estimated standard error, and CP is the coverage probability based on a 95% nominal confidence interval.

2.4.3 Computational Advantage

Both the pairwise likelihood and semiparametric approach reduce the computational burden involved with evaluation of the integral L_i. The pairwise likelihood approach for one replication of this simulation study takes about 12 minutes, while the semiparametric approach takes about 30 seconds. For small datasets, this 25-fold improvement may not be significant, but for datasets as large as or larger than our motivating insurance data, this improvement has a significant impact on feasibility of computation. The full likelihood approach takes about 1 hour for one replication. SAS is used for estimation of the likelihood, pairwise likelihood and semiparametric approaches.[7] Procedures were run on a Windows Server 2008 R2 Datacenter with an Intel Xeon CPU, 4 2.92 GHz processors and 128GB of RAM.

2.5 An Empirical Application: Insurance Data

2.5.1 Description of Insurance Data

The semiparametric and pairwise likelihood methods are used to analyze the insurance claim count data of matched records for home and auto insurance for each policyholder over the course of nine years, 1998 – 2006. At the beginning of each claim year, a snapshot of policy and policy holder characteristics is observed that is linked to the number of three types of insurance claims - home all perils, auto collision and auto com-

[7]Programs from the authors of the pairwise approach are used in this research (Fieuws and Verbeke, 2006). The implementation of their method uses the "NLMIXED" procedure in SAS which directly maximizes an approximate integrated likelihood using Gauss-Hermite quadrature and quasi-Newton optimization.

prehensive - filed during the course of the year. See Section 2 for more details. The multiplicative random effects Poisson model corresponds to the following set of mixed models:

$$\begin{cases} E(\mathbf{y}_{i1}|\mathbf{x}_{i1}, u_{i1}) = u_{i1} exp\left(\mathbf{x}_{i1}^T \beta_1 + \text{offset}_{i1}\right) \\ E(\mathbf{y}_{i2}|\mathbf{x}_{i2}, u_{i2}) = u_{i2} exp\left(\mathbf{x}_{i2}^T \beta_2 + \text{offset}_{i2}\right) \\ E(\mathbf{y}_{i3}|\mathbf{x}_{i3}, u_{i3}) = u_{i3} exp\left(\mathbf{x}_{i3}^T \beta_3 + \text{offset}_{i3}\right) \end{cases}$$

where $k = 1, 2, 3$ corresponds to home, collision and comprehensive claim counts, respectively. In this case, the vector of random effects $\mathbf{u}_i$ is three dimensional with covariance matrix Σ. The set of 35 policy and household characteristics related to home claims, $\mathbf{x}_{i1}$, include credit score, home value, home age, number of families, alarm/protection device indicators, home use, occupancy status, construction type, fire protection classes and consolidated territory codes. The set of 37 policy and household characteristics related to auto claims, $\mathbf{x}_{i2}$ and $\mathbf{x}_{i3}$, include credit score, number of drivers, number of vehicles, driver age, driver marital status, vehicle age, vehicle use, vehicle safety features and consolidated territory codes.

2.5.2 Empirical Results

The research question of interest dictates the need for joint modeling of the multivariate longitudinal outcomes. To account for association across time and outcomes the semiparametric and pairwise likelihood approaches are implemented. For comparison purposes, the results from a univariate Poisson-lognormal generalized linear mixed model are also presented. The results for the association parameters are presented in Table 2.4.

Table 2.4: Association Parameter Results for Analysis of Insurance Claim Data

	Balanced Panel 8,731 Policies, 78,579 Observations						Unbalanced Panel 62,425 Policies, 294,917 Observations			
Parameter	Semiparametric		Pairwise		Uni. GLMM		Semiparametric		Uni. GLMM	
σ_{11}	0.467	(0.072)	0.396	(0.035)	0.395	(0.036)	0.320	(0.077)	0.531	(0.025)
σ_{22}	0.155	(0.036)	0.181	(0.021)	0.181	(0.022)	0.123	(0.023)	0.204	(0.013)
σ_{33}	0.551	(0.182)	0.540	(0.077)	0.537	(0.080)	0.571	(0.109)	0.745	(0.055)
σ_{12}	0.136	(0.021)	0.137	(0.019)	.	(.)	0.078	(0.014)	.	(.)
σ_{13}	0.224	(0.038)	0.223	(0.035)	.	(.)	0.227	(0.024)	.	(.)
σ_{23}	0.104	(0.029)	0.120	(0.027)	.	(.)	0.122	(0.019)	.	(.)
ρ_{12}	0.504	(0.104)	0.513	(0.079)	.	(.)	0.392	(0.091)	.	(.)
ρ_{13}	0.442	(0.110)	0.481	(0.086)	.	(.)	0.531	(0.100)	.	(.)
ρ_{23}	0.355	(0.123)	0.384	(0.092)	.	(.)	0.459	(0.093)	.	(.)

Notes: Standard errors are in parentheses. The balanced panel sample includes those customers with both home and auto policies in force for all 9 years. The unbalanced panel sample includes those customers with both home and auto policies in force at any time in the 9 years.

In the analysis of the balanced panel subset of the insurance data, the semiparametric and pairwise likelihood approaches lead to similar conclusions regarding the association between unobserved heterogeneity, though inferential conclusions about the correlation parameters are affected by the fact that the standard errors for the variance estimates are larger in the semiparametric approach, consistent with the simulation study results. If the multivariate structure of the count outcomes is ignored, the GLMM based on the lognormality specification for the random effects finds close to identical estimates and standard errors to those obtained from the pairwise likelihood method. In contrast, in the analysis of the more complete unbalanced panel sample of the insurance data, the semiparametric and GLMM methods lead to significant differences in the estimates of the variance parameters. Note that it is computationally feasible to estimate the association parameters in multivariate longitudinal claim count model with the semiparametric approach, but not the pairwise likelihood approach. While the pairwise likelihood is computationally

prohibitive for the unbalanced sample, the GLMM and simulation study results suggest that the lognormality assumption may result in an overestimation of the association parameters. The semiparametric method is robust to this potential bias.

These results imply an interesting economic result. Wald-type tests based on the standard errors obtained via the Delta method indicate that all three pairwise correlations are statistically significantly positive regardless of the approach used or sample analyzed. The relation between the unobserved heterogeneity in all of the pairwise claim rates is positive, but it is most positive between home and collision in the balanced panel subsample and most positive between home and comprehensive in the unbalanced panel sample. The magnitude of the variance parameters estimates of the unobserved heterogeneity indicates that there is more variation in the unobserved effect for home and comprehensive coverage than collision coverage in the balanced panel subsample, .467/.551 and .155 respectively, and more variation in the unobserved effect for comprehensive coverage than home and collision coverage in the unbalanced panel sample, .571 and .320/.123, respectively. With respect to the regression parameters, $(\beta_1, \beta_2, \beta_3)$, there is little variation in both the point estimates and the standard errors between the semiparametric, pairwise likelihood and univariate approaches. For results concerning the association and regression parameters and their economic interpretation, please refer to Chapter 3 (Barseghyan et al., 2012).

With the semiparametric approach for multivariate longitudinal count data, estimation of the association of the unobserved heterogeneity is computationally feasible for large datasets with a large number of covariates and robust to distributional specifications of the latent effects that are inherently untestable. The analysis of the richer unbalanced data suggests that the random effect distributional assumption may have significant impact on the conclusions about the relation between unobserved policy-specific

time-invariant characteristic across different types of insurance coverage.

2.5.3 Computational Advantage

Just as in the simulation study, the semiparametric approach requires much less computing time than the pairwise likelihood approach, about 2.5 hours versus about 60 hours for the balanced case and about 22 hours for semiparametric estimation in the unbalanced case, while the full likelihood approach and the pairwise likelihood approach for the unbalanced panel sample is computationally prohibitive. The assessment of the association of unobserved heterogeneity in the insurance data with 3 count outcomes, a maximum of 9 time periods, $62,425$ policies and $294,917$ total observations is computationally feasible with the semiparametric approach, but prohibitively computationally intense with the pairwise likelihood method. Both methods are programmed in SAS using the NLMIXED procedure and SAS/IML. The empirical analysis was run on a Windows Server 2008 R2 Enterprise with an Intel Xeon CPU, 4 2.93 GHz processors and 128GB of RAM.

2.6 Discussion

2.6.1 Complications of Sparse Counts

The descriptive statistics in Table 2.1 illustrate the abundance of zeros observed in the sample of insurance data we analyze with the semiparametric and pairwise likelihood methods. The methods validly account for zeros through the specification of the mean, but some complications arise due to the sparse nature of the response. Overall, the

semiparametric approach performs well, however the advantage of robustness is compromised as the simulation study shows that the pairwise likelihood approach is similarly robust to distributional misspecification when the data exhibits a combination of smaller sample size and sparse counts. This combination also results in the inflation of the variance estimates of the latent effect variance parameters. In the simulation studies, we find that the estimate of the standard error for the variance parameters sometimes exceed twice the true sampling variability. This result is reflected in the application where the standard errors for the variance parameters using the semiparametric approach are much larger than those obtained from the pairwise likelihood approach. This overestimation seems to occur because of the sparse nature of the data, as the simulation study with non-sparse counts show that the estimator of the standard error does not exhibit this property.

For practical purposes, these results indicate that the researcher should seriously take into account the dimension of the data and the sparsity of the count outcome when making inference regarding the association parameters. For example, in a simple univariate, constant mean model, Figure 2.3 depicts the relation between the number of time periods, number of subjects and sparsity of positive counts as measured through the mean parameters with respect to RMSE. There clearly is a tradeoff between data dimension and sparse counts that the empirical researcher must accommodate in inferential conclusions. Furthermore, while the computational advantage of the semiparametric method remains, the robustness property is compromised in the sparse count and small sample settings.

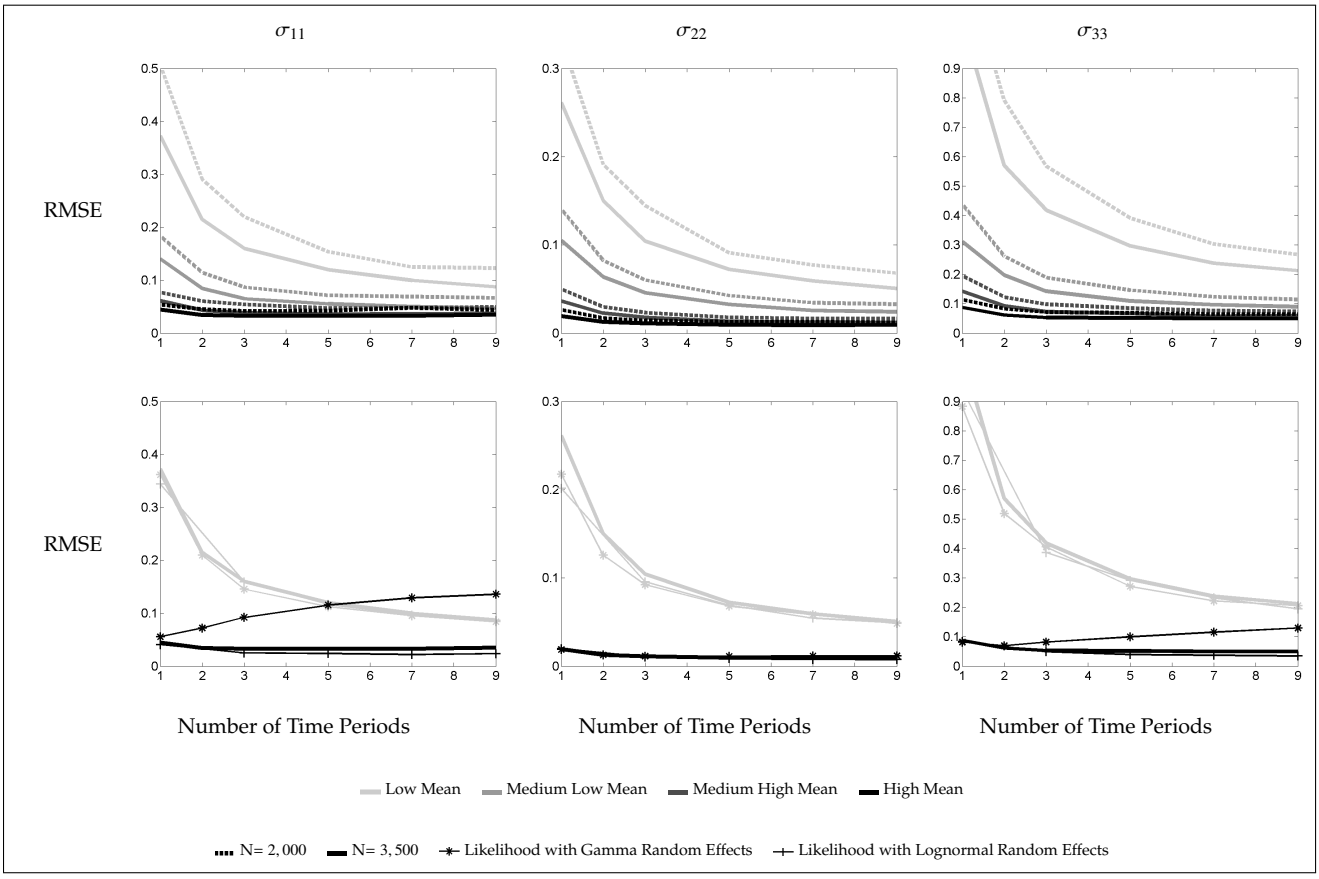

Figure 2.3: Semiparametric Approach RMSE of Variance Parameter Estimates from Simple Simulation Study

2.6.2 Informative Missingness

In the insurance data application, the sample for analysis is restricted to those policies that have both home and auto policies in the nine year period with valid values for all variables used in the analysis. In this unbalanced panel setting, the consistency results for the semiparametric approach are valid by implicitly assuming the dropout/missing data process is MCAR, i.e. the cancellation/non-renewal process is not related to the claim count process. For example, the "riskier" policies are not any more or less likely

to be canceled or not renewed. If the missing data process is MAR or NMAR then the semiparametric approach fails to provide consistent estimates. Adjustments, such as joint models of counts and dropouts or inverse probability weighting (Robins et al., 1995; Cook and Li, 2002; Diggle et al., 2002), have not been implemented here. The relation between claim rates and cancellation/non-renewal decisions is explored in Chapter 4.

2.7 Conclusion

We propose a semiparametric method for multivariate longitudinal data based on a random effects model and GEE. Joint modeling of a multivariate outcome measured over time allows for estimation and inference on association parameters for unobserved heterogeneity. The semiparametric method shows an improvement over the pseudo-likelihood based pairwise method when the random effect distribution is misspecified. In addition, the semiparametric method avoids intractable high-dimensional integration, resulting in considerable computational advantages: about a 25-fold decrease in computing time compared to the pairwise likelihood approach.

In the case of the insurance data, the association of the unobserved effects is of economic interest as it provides insight into underlying risk related characteristics of the policy holders. Application of the semiparametric method to the insurance claim data takes advantage of the joint information that multivariate longitudinal data contains about subject specific heterogeneity between count outcomes, which would otherwise be computationally prohibitive with likelihood methods. We find strong correlation in the unobserved heterogeneity of the claim rate model between all pairwise combinations of insurance coverage types: home, auto collision and auto comprehensive.

CHAPTER 3

UNOBSERVED HETEROGENEITY IN INSURANCE CLAIMS

Joint work with Levon Barseghyan, Francesca Molinari and Joshua Teitelbaum

3.1 Introduction

Unobserved heterogeneity in claim risk in theory leads to adverse selection in insurance markets. Legal restrictions on experience rating preclude insurers from meliorating any welfare loss from adverse selection.[1] In the United States, the laws of most (if not all) states limit the ability of insurance companies to engage in experience rating within and across different lines of property insurance. This research explores the statistical and economic advantages that cross-coverage information provides in claim risk models, suggesting that legal restrictions on experience rating exacerbate any dead weight loss from adverse selection.

We demonstrate the economic significance of the within-coverage variances and cross-coverage correlations of unobserved heterogeneity with respect to three lines of insurance coverage: auto collision (c), auto comprehensive (m), and home (h). Using the semiparametric, moment-based approach of Chapter 2, we estimate the variance-covariance matrix of unobserved heterogeneity, $\widehat{\mathbf{\Sigma}}$, associated with the correlated random effects Poisson model. We find a strong positive correlation, $\widehat{\rho} = (0.663, 0.293, 0.559)$, between unobserved heterogeneity in all pairwise combinations of insurance coverages, indicating potential

[1] Under experience rating, an insured's premium is adjusted or modified based on his or her actual loss experience. Experience rating is not to be confused with classification rating, under which an insured's premium reflects the collective loss experience of all insureds in the insured's risk class (which class may be defined in part by actual loss experience).

benefits in taking advantage of cross-coverage information.

The estimated strong positive correlation and within-coverage levels of variance of unobserved heterogeneity, obtained from an unbalanced panel dataset of $62,425$ households who held all three coverages over the course of a nine year period, are used to demonstrate the economic significance of information on variance-covariance structure of unobserved heterogeneity, including incremental and independent value of information on cross-coverage correlation structure of unobserved heterogeneity. We show that utilizing the information in $\widehat{\boldsymbol{\Sigma}}$ leads to material refinements of the predicted claim rates in the tricoverage sample. There is significant incremental value of utilizing the information on the cross-coverage correlation structure of unobserved heterogeneity, in addition to the information on the within-coverage variance of unobserved heterogeneity. Simulation studies illustrate the value of the utilizing the information in the variance-covariance matrix, and investigate the independent value of the information on the cross-coverage correlation structure, in terms of (i) improving the accuracy of the predicted claim rates and (ii) updating a household's predicted claim rates to reflect subsequent claims experience.

The rest of this chapter is organized as follows: Section 3.2 describes the motivating insurance data; Section 3.3 summarizes the model and estimation strategy; Section 3.4 presents the empirical results; Section 3.5 evaluates the economic significance of incorporating unobserved heterogeneity through empirical findings and simulation studies; and Section 3.6 concludes.

3.2 Description of the Data

The source of the data is a large property and casualty insurance company. The company offers several lines of insurance, including auto and home. The full data set includes annual information on more than $400,000$ households who held auto or home policies between 1998 and 2006. The data contain all the information in the company's records regarding the households and their policies (premiums, deductibles, etc.). In addition, the data record the number of claims that each household filed with the company under each of its policies during the period of observation.

We focus our attention on three lines of coverage: auto collision, auto comprehensive, and home all perils. Auto collision coverage pays for damage to the insured vehicle caused by a collision with another vehicle or object, without regard to fault. Auto comprehensive coverage pays for damage to the insured vehicle from all other causes (e.g., theft, fire, flood, windstorm, glass breakage, vandalism, hitting or being hit by an animal or by falling or flying objects), without regard to fault. Home all perils coverage pays for damage to the insured home from all causes (e.g., fire, windstorm, hail, tornadoes, vandalism, or smoke damage), except those that are specifically excluded (e.g., flood, earthquake, or war).[2]

In most of the analysis, we consider an unbalanced panel of $62,425$ households who held all three coverages (auto collision, auto comprehensive, and home) in one or more years between 1998 and 2006. In all, this tricoverage sample comprises $294,917$ household-years. Descriptive statistics are set forth in Appendix B.

[2]For simplicity, we often refer to home all perils simply as home.

Table 3.1: Summary of Claims, Premiums, and Deductibles

Tricoverage Sample (294,917 household-years)

	Mean	Std. Dev.	Minimum	Maximum
Claims (count):				
Collision	.107	.334	0	5
Comprehensive	.032	.188	0	5
Home	.079	.299	0	6
Premiums (dollars):				
Collision	200	103	20	2,520
Comprehensive	127	70	6	2,524
Home	548	309	50	10,224
Deductible (dollars):				
Collision	396	181	100	1,000
Comprehensive	273	176	50	1,000
Home	350	242	100	5,000

Table 3.1 summarizes the claims, premiums, and deductibles in the tricoverage sample. Additional details are set forth in Appendix B. The mean number of claims per household-year is 0.107 in auto collision, 0.032 in auto comprehensive, and 0.079 in home. On average, households paid annual premiums of $200 in auto collision, $127 in auto comprehensive, and $548 in home. The mean deductibles per household-year are $396, $273, and $350 in auto collision, auto comprehensive, and home, respectively. The modal deductibles are $500 in auto collision, $200 in auto comprehensive, and $250 in home.

3.3 Model and Estimation Strategy

3.3.1 Model

A standard regression model for longitudinal univariate count data is the Poisson random effects model. We extend this model to multivariate count data—here, claim counts under three types of insurance coverage—by allowing for correlated random effects.

Let y_{itk} denote the number of claims for household i in year t under coverage k, where $i = 1, ..., N$, $t = 1, ..., T_i$, and $k \in \{c, m, h\}$.[3] Similarly, let $\mathbf{x}_{itk}$ denote a vector of observables (plus a constant) for household i in year t under coverage k. Let λ_{itk} denote household i's baseline claim rate in year t under coverage k, and let u_{ik} denote a time-constant random effect for household i under coverage k. Both λ_{itk} and u_{ik} are unobserved.

We assume

$$\mathbf{y}_{itk}|\mathbf{x}_{itk} \sim Poisson\left(\lambda_{itk}u_{ik}\right),$$

where

$$\lambda_{itk} = \exp\left(\mathbf{x}'_{itk}\beta_k\right)$$

and

$$\mathbf{u}_i \equiv \begin{bmatrix} u_{ic} \\ u_{im} \\ u_{ih} \end{bmatrix} \overset{iid}{\sim} \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \sigma_c^2 & \rho_{mc}\sigma_m\sigma_c & \rho_{hc}\sigma_h\sigma_c \\ \rho_{cm}\sigma_c\sigma_m & \sigma_m^2 & \rho_{hm}\sigma_h\sigma_m \\ \rho_{ch}\sigma_c\sigma_h & \rho_{mh}\sigma_m\sigma_h & \sigma_h^2 \end{bmatrix} \right).$$

[3] In the set of coverages, c denotes auto collision, m denotes auto comprehensive, and h denotes home.

The parameters to be estimated are

$$\beta \equiv \begin{bmatrix} \beta_c \\ \beta_m \\ \beta_m \end{bmatrix} \text{ and } \boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_c^2 & \rho_{mc}\sigma_m\sigma_c & \rho_{hc}\sigma_h\sigma_c \\ \rho_{cm}\sigma_c\sigma_m & \sigma_m^2 & \rho_{hm}\sigma_h\sigma_m \\ \rho_{ch}\sigma_c\sigma_h & \rho_{mh}\sigma_m\sigma_h & \sigma_h^2 \end{bmatrix}.$$

Of principal interest is the variance-covariance matrix, $\boldsymbol{\Sigma}$, which captures both the within-coverage variance of unobserved heterogeneity, $\sigma^{\mathbf{2}} \equiv (\sigma_c^2, \sigma_m^2, \sigma_h^2)$, and the cross-coverage correlation structure of unobserved heterogeneity, $\rho \equiv (\rho_c, \rho_m, \rho_h)$.

3.3.2 Estimation Strategy

The marginal likelihood function may be written as

$$L_i = \int_{u_{ih}} \int_{u_{im}} \int_{u_{ic}} \left\{ \prod_k \prod_t \exp\left(-u_{ik}\lambda_{itk}\right) \frac{(u_{ik}\lambda_{itk})^{y_{itk}}}{y_{itk}!} \right\} f(u_{ic}, u_{im}, u_{ih}) du_{ic} du_{im} du_{ih}$$

where $f(u_{ic}, u_{im}, u_{ih})$ is the trivariate density of $\mathbf{u}_i$. A standard parametric approach is to specify f and estimate the model by maximum likelihood. Typical specifications of f include the lognormal distribution and the gamma distribution (in which case L_i reduces to the product of negative binomial densities). In our case, however, the standard approach has two drawbacks. First, maximizing the likelihood function may be computationally intractable. The likelihood function not only involves a multidimensional integral, but, depending on f, it also may not have a closed-form expression. Second, if f is incorrectly specified, the maximum likelihood estimator may be inconsistent.

We adopt the semiparametric, moments-based approach of Chapter 2 (Morris, 2011), which provides a computationally tractable method for consistent estimation of β and $\boldsymbol{\Sigma}$ for all possible densities f. Under this approach, estimation is via generalized estimating equations (GEE) based on marginal moments. Given the assumptions of our model,

we can derive the first and second marginal moments and use them to construct estimating equations for β and $\mathbf{\Sigma}$. More specifically, we use the first marginal moment to define a quasi-score equation, where the associated estimating equation for β is based on a weighted least squares estimator with the weight matrix defined by the covariance structure derived from the second marginal moment. The estimating equation for $\mathbf{\Sigma}$ is based on the relation between the empirical variance estimate and the model defined covariance structure. The two estimating equations are solved iteratively to obtain $\widehat{\beta}$ and $\widehat{\mathbf{\Sigma}}$.[4]

3.4 Estimation Results

3.4.1 Regression Estimates

Table 3.2 presents the estimates of the association parameters, σ^2 and ρ, implied by $\widehat{\mathbf{\Sigma}}$. The estimates suggest that the variance of unobserved heterogeneity is lowest in auto collision ($\widehat{\sigma}^2_c = 0.11$), and is roughly four times higher in auto comprehensive ($\widehat{\sigma}^2_m = 0.40$) and home ($\widehat{\sigma}^2_h = 0.41$). More importantly, the estimates reveal that unobserved heterogeneity is correlated across coverages, and that each pairwise correlation is positive and statistically significant. Perhaps not surprisingly, the strongest correlation is between auto collision and auto comprehensive ($\widehat{\rho}_{cm} = 0.66$). More surprising, however, is the fairly strong correlation between auto comprehensive and home ($\widehat{\rho}_{mh} = 0.56$). After all, one might rea-

[4]For further details, see Chapter 2 (Morris, 2011). This approach is an extension of quasi-generalized pseudo maximum likelihood (QGPML) estimators developed by Gourieroux et al. (1984b,a) and the extended GEE approach developed by Prentice (1988). The QGPML method can be characterized as first order GEE with a specific association structure. Prentice (1988) introduces an extension of first order GEE that utilizes a second set of estimating equations to jointly estimate the association parameters. QGPML can be embedded in the GEE framework resulting in commonly studied consistency and asymptotic results for simultaneous inference on both the regression parameters and the association parameters.

Table 3.2: Association Parameter Estimates

Tricoverage Sample (294,917 household-years)

	Est.	SE
Variances:		
Collision	.107*	.021
Comprehensive	.399*	.091
Home	.405*	.011
Covariances:		
Collision and Comprehensive	.137*	.018
Collision and Home	.061*	.020
Comprehensive and Home	.225*	.024
Correlations:		
Collision and Comprehensive	.663*	.135
Collision and Home	.293*	.099
Comprehensive and Home	.559*	.087

* Significant at 5 percent level.

sonably conjecture that claim risk in auto comprehensive and home reflect force majeure risk more than household specific risk. The weakest correlation is between auto collision and home ($\widehat{\rho}_{ch} = 0.29$). Even this correlation, however, is economically significant, as we demonstrate below in Section 3.5.

The estimates of the regression parameters, β, are reported in Appendix B. Although β is not the object of principal interest, the estimates reveal several noteworthy facts. First, auto claim rates (collision and comprehensive) are negatively related to insurance score but positively related to the age and number of vehicles.[5] However, they are not correlated with vehicle safety features (passive restraint, anti-theft, and anti-lock brakes). Second, collision claim rates are negatively related to the age of the primary driver and are higher for households in which the primary driver is female. Conversely, comprehensive

[5]Insurance score is based on information contained in credit reports.

claim rates are positively related to the age of the primary driver and are lower for households in which the primary driver is female. Third, collision claim rates are higher for households with three or more drivers. Finally, home claim rates are negatively related to insurance score but positively related to the age and insured value of the home. In addition, they are higher for homes that are used for farming or business and for homes that are not the owner's primary residence. Home claim rates, however, are not correlated with home safety features (masonry construction, distance to fire hydrant, and alarm or other protection).

3.4.2 Sensitivity Checks

As checks of the sensitivity of the association parameter estimates, we re-estimate the model on two alternative samples and also on a number of subsamples of the tricoverage sample.[6]

The two alternative samples we consider are: (A) a balanced panel of $8,731$ households ($78,579$ household-years) who held all three coverages (auto collision, auto comprehensive, and home); and (B) an unbalanced panel of $203,731$ households ($1,019,170$ household-years) who held both auto coverages (collision and comprehensive). Table 3.3 reports the association parameter estimates for both alternative samples. They are largely consistent with the estimates for the tricoverage sample. The only difference is that the correlation between auto collision and home is higher in the balanced panel (alternative sample A) than it is in the tricoverage sample. In the balanced panel, this correlation is as

[6]To ease the computational burden, the sensitivity analysis uses GLM estimates of the regression parameters (assuming the random effects follow a lognormal distribution). In the tricoverage sample, the semiparametric and GLM estimates for β are nearly identical ($R^2 = 0.9998$). Thus, we are confident that using the GLM estimates for β does not corrupt the sensitivity analysis of the semiparametric estimates of the association parameters.

Table 3.3: Association Parameter Estimates - Alternative Samples

	Tri. Sample 62,425 households 294,917 obs		Alt. Sample A 8,731 households 78,579 obs		Alt. Sample B 203,731 households 1,019,170 obs	
	Est.	SE	Est.	SE	Est.	SE
Variances:						
Collision	.107*	.021	.114*	.033	.093*	.012
Comprehensive	.399*	.091	.342*	.140	.402*	.052
Home	.405*	.011	.401*	.072	.	.
Covariances:						
Collision and Comprehensive	.137*	.018	.123*	.030	.131*	.010
Collision and Home	.061*	.020	.121*	.020	.	.
Comprehensive and Home	.225*	.024	.209*	.038	.	.
Correlations:						
Collision and Comprehensive	.663*	.135	.622*	.218	.680*	.081
Collision and Home	.293*	.099	.564*	.136	.	.
Comprehensive and Home	.559*	.087	.563*	.161	.	.

* Significant at 5 percent level.

Notes: The tricoverage sample comprises an unbalanced panel of households who held all three coverages (auto collision, auto comprehensive, and home) in one or more years between 1998 and 2006. Alternative sample A comprises a balanced panel of households who held all three coverages (auto collision, auto comprehensive, and home). Alternative sample B comprises an unbalanced panel of households who held both auto coverages (collision and comprehensive).

high as the correlation between auto comprehensive and home, whereas in the tricoverage sample it was roughly has as high.

The subsamples of the tricoverage sample we consider are: households with low and high insurance scores; households with low and high home values; households with young and old primary drivers; households with female and male primary divers; and households with married primary drivers. In each case, the subsample is defined by household characteristics at the time of first observation. For continuous variables, low and high (or, in the case of age, young and old) are defined as the bottom third and top third, respectively. The association parameter estimates for these subsamples are reported

in Appendix B. They also are largely consistent with the estimates for the tricoverage sample.

3.4.3 Moral Hazard

Our approach implicitly assumes that a household's claim risk is not influenced by its deductible choice. That is, we assume households do not suffer from moral hazard. In particular, we assume there is neither ex ante moral hazard (deductible choice does not influence the frequency of claimable events) nor ex post moral hazard (deductible choice does not influence the decision to file a claim). The empirical evidence on moral hazard in auto insurance markets is mixed (Cohen and Siegelman, 2010), and we are not aware of any empirical evidence on moral hazard in home insurance markets. Because the choice of deductible usually has a small effect on the overall level of coverage, it seems reasonable to assume there is no ex ante moral hazard. However, because the damage from a claimable event occasionally may be less than the chosen deductible, it may be less reasonable to assume there is no ex post moral hazard. As a check of the sensitivity of the association parameter estimates to our assumption on moral hazard, we re-estimate the model separately for "low deductible" and "high deductible" households. We define a household as a "low deductible" household if none of its deductibles is greater than \$250. Conversely, we define a household as a "high deductible" household if at least one of its deductibles is greater than \$250.[7] Table 3.4 reports the association parameter estimates for low and high deductible households.[8] They are largely consistent with the estimates for the tricoverage sample, suggesting that moral hazard is not an issue.

[7] Recall that each household in the tricoverage sample has three deductibles, one for auto collision, one for auto comprehensive, and one for home.

[8] As before, the re-estimations use GLM estimates of the regression parameters (assuming the random effects follow a lognormal distribution).

Table 3.4: Association Parameter Estimates - Low and High Deductible Households

	Tri. Sample 62,425 households 294,917 obs		Deductible ≤ $250 22,072 households 120,213 obs		Deductible > $250 40,353 households 174,704 obs	
	Est.	SE	Est.	SE	Est.	SE
Variances:						
Collision	.107*	.021	.094*	.028	.108*	.029
Comprehensive	.399*	.091	.337*	.128	.450*	.127
Home	.405*	.011	.388*	.055	.246*	.106
Covariances:						
Collision and Comprehensive	.137*	.018	.138*	.027	.129*	.025
Collision and Home	.061*	.020	.088*	.017	.058*	.019
Comprehensive and Home	.225*	.024	.224*	.034	.217*	.033
Correlations:						
Collision and Comprehensive	.663*	.135	.776*	.243	.586*	.160
Collision and Home	.293*	.099	.458*	.116	.352*	.146
Comprehensive and Home	.559*	.087	.619*	.157	.652*	.195

* Significant at 5 percent level.

3.4.4 Excess Zeros

Excess zeros are a common problem when using a Poisson model for count data. Table 3.5 compares for each coverage the empirical distribution of claim counts in the tricoverage sample with the predicted distribution of claim counts, when the latter is calculated ignoring random effects and using predicted baseline claim rates, $\widehat{\lambda}_{itk} \equiv \exp(\mathbf{x}'_{itk}\widehat{\beta}_k)$. We refer to $\widehat{\lambda}_{itk}$ as the prior claim rate. Table 3.5 suggests that, even without random effects, we do not have an excess zeros problem. Although the model underpredicts the percentage of zeros, it does so by less than one half of one percentage point (in absolute terms), or less than 0.5 percent (in percentage terms). Moreover, the model overpredicts the percentage of ones and underpredicts the percentage twos and threes. This suggests that there is room to improve the fit of the model with respect to zero and non-zero counts. In

Table 3.5: Distribution of Claim Counts - Actual versus Predicted

Tricoverage Sample (294,917 household-years)

	Collision		Comprehensive		Home	
Count	Actual	Predicted	Actual	Predicted	Actual	Predicted
0	90.09	89.95	96.95	96.73	92.90	92.48
1	9.22	9.45	2.88	3.09	6.40	7.17
2	0.64	0.57	0.16	0.07	0.63	0.33
3	0.05	0.03	0.01	-	0.05	0.01
4	-	-	-	-	-	-

Note: Values are percentages. Predicted distributions are based on prior claim rates. Dash indicates less than 0.01 percent.

the next section, we demonstrate that we can materially refine the predicted claim rates and improve the model's overall fit by including correlated random effects to account for unobserved heterogeneity.

3.5 Economic Significance of Unobserved Heterogeneity

In this section, we demonstrate the value of the information about the within-coverage variance of unobserved heterogeneity (σ^2) and the cross-coverage correlation structure of unobserved heterogeneity (ρ) contained in the estimated variance-covariance matrix, $\widehat{\mathbf{\Sigma}}$. We first show that utilizing the information in $\widehat{\mathbf{\Sigma}}$ leads to material refinements of the predicted claim rates in the tricoverage sample. We also show the incremental value of utilizing the information on the cross-coverage correlation structure of unobserved heterogeneity, in addition to the information on the within-coverage variance of unobserved heterogeneity. We then report the results of a simulation study, which further investigates the value of the utilizing the information in the variance-covariance matrix in terms

of (i) improving the accuracy of the predicted claim rates and (ii) updating a household's predicted claim rates to reflect subsequent claims experience.

3.5.1 Tricoverage Sample

In order to demonstrate that utilizing the information in $\widehat{\boldsymbol{\Sigma}}$ leads to material refinements of the predicted claim rates in the tricoverage sample, we compare the empirical distribution of the prior claim rate $\widehat{\lambda}_{itk}$ with that of the multivariate posterior claim rate $\widehat{\theta}_{itk} \equiv \widehat{\lambda}_{itk} E^{MV}(u_{ik}|\mathbf{y}_i)$, where $\mathbf{y}_i = (y_{itc}, ..., y_{iT_ic}, y_{itm}, ..., y_{iT_im}, y_{ith}, ..., y_{iT_ih})$ and $E^{MV}(u_{ik}|\mathbf{y}_i)$ is calculated assuming $[u_{ic}\ u_{im}\ u_{ih}]' \overset{iid}{\sim} lognormal([1\ 1\ 1]', \widehat{\boldsymbol{\Sigma}})$. Figure 3.1 plots, for each coverage $k = l, m, h$, the kernel density of $\eta_{itk} \equiv (\widehat{\theta}_{itk} - \widehat{\lambda}_{itk})/\widehat{\lambda}_{itk}$. Further details are set forth in Table 3.6. For households with negative values of η_{itk}, the mean value of η_{itk} is −7 percent in auto collision, −13 percent in auto comprehensive, and −14 percent in home. For a quarter of these households, η_{itk} is less than −9 percent in auto collision, −19 percent in auto comprehensive, and −20 percent in home. For a tenth of these households, η_{itk} is less than −12 percent in auto collision and −24 percent in both auto comprehensive and home. The numbers are even more striking for households with positive values of η_{itk}. For these households, the mean value of η_{itk} is +10 percent in auto collision, +23 percent in auto comprehensive, and +28 percent in home. For a quarter of these households, η_{itk} exceeds +14 percent in auto collision, +31 percent in auto comprehensive, and +37 percent in home. For a tenth of these households, η_{itk} exceeds +23 percent in auto collision, +53 percent in auto comprehensive, and +65 percent in home. The numbers are remarkably similar for households with low, medium, and high prior claim rates,[9] suggesting that the

[9] A prior claim rate is "low" if it is in the bottom quartile and "high" if it is in the top quartile. It is "medium" otherwise. The respective low and high cutoffs are 0.078 and 0.127 in auto collision, 0.016 and 0.044 in auto comprehensive, and 0.054 and 0.096 in home.

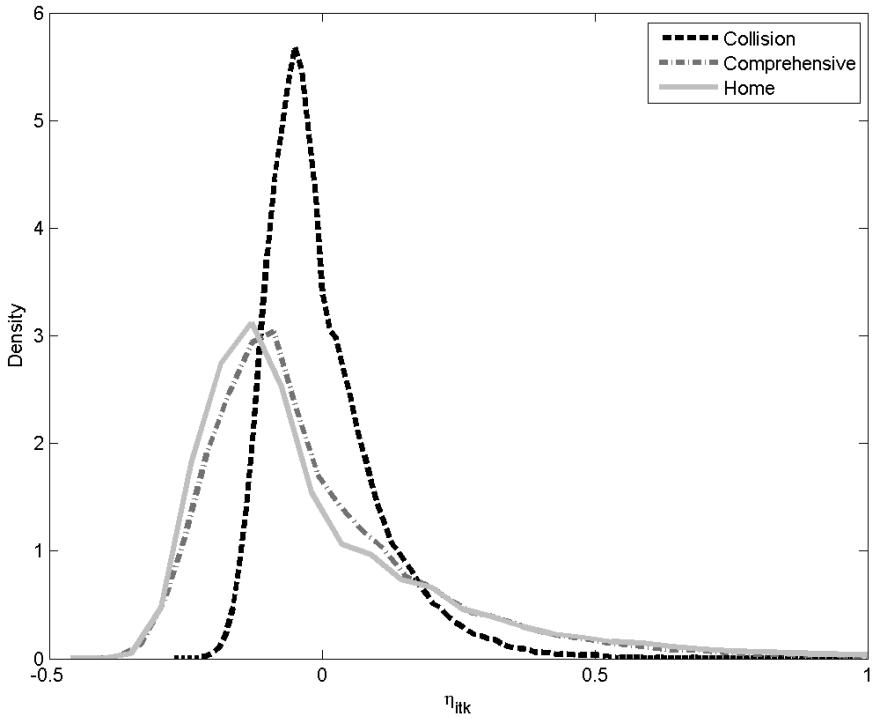

Figure 3.1: Kernel Density of η_{itk}

value of the information in $\widehat{\mathbf{\Sigma}}$ is robust to differences in baseline claim risk.

To show the incremental value of utilizing the information on the cross-coverage correlation structure of unobserved heterogeneity, as opposed to utilizing only the information on within-covariance variance, we compare the empirical distribution of $\widehat{\theta}_{itk}$ with that of the univariate posterior claim rate $\widehat{\vartheta}_{itk} \equiv \widehat{\lambda}_{itk} E^{UV}(u_{ik}|\mathbf{y}_i)$, where $E^{UV}(u_{ik}|\mathbf{y}_i)$ is calculated assuming $u_{ik} \overset{iid}{\sim} lognormal(1, \widehat{\sigma}_k^2)$ for $k = l, m, h$. Figure 3.2 plots, for each coverage k, the kernel density of $\zeta_{itk} \equiv (\widehat{\theta}_{itk} - \widehat{\vartheta}_{itk})/\widehat{\vartheta}_{itk}$. Further details are set forth in Table 3.7. For households with negative values of ζ_{itk}, the mean value of ζ_{itk} is -3 percent in auto collision, -10 percent in auto comprehensive, and -4 percent in home, and for a tenth of these households ζ_{itk} is less than -6 percent in auto collision, -17 percent in auto comprehensive, and

Table 3.6: Descriptive Statistics for $\eta = \frac{\theta - \lambda}{\lambda}$

	$\eta < 0$				$\eta > 0$			
Prior Claim Rates	Obs.	Mean	10^{th} Pctile	25^{th} Pctile	N	Mean	75^{th} Pctile	90^{th} Pctile
				Collision				
All	180,909	-0.066	-0.121	-0.093	114,008	0.101	0.140	0.227
Low	46,914	-0.055	-0.100	-0.080	26,815	0.094	0.129	0.213
Medium	89,988	-0.066	-0.120	-0.094	57,471	0.100	0.139	0.223
High	44,007	-0.078	-0.143	-0.110	29,722	0.110	0.152	0.249
				Comprehensive				
All	188,792	-0.132	-0.236	-0.186	106,125	0.231	0.310	0.531
Low	47,384	-0.111	-0.198	-0.157	26,345	0.198	0.270	0.455
Medium	94,742	-0.131	-0.233	-0.186	52,717	0.231	0.307	0.533
High	46,666	-0.155	-0.273	-0.220	27,063	0.264	0.353	0.593
				Home				
All	196,205	-0.142	-0.241	-0.198	98,712	0.280	0.367	0.646
Low	51,136	-0.120	-0.207	-0.166	22,593	0.273	0.350	0.630
Medium	97,208	-0.146	-0.238	-0.200	50,251	0.277	0.365	0.642
High	47,861	-0.159	-0.271	-0.227	25,868	0.292	0.390	0.663

−8 percent in home. Again, the numbers are more striking for households with positive values of η_{itk}. For these households, the mean value of ζ_{itk} is +7 percent in auto collision, +16 percent in auto comprehensive, and +9 percent in home, and for a tenth of these households ζ_{itk} exceeds +15 percent in auto collision, +36 percent in auto comprehensive, and +21 percent in home. As before, the numbers are very similar for households with low, medium, and high prior claim rates, suggesting that the incremental value of the information in $\widehat{\rho}$ is robust to differences in baseline claim risk.

Table 3.8 reveals that utilizing the information on unobserved heterogeneity contained in $\widehat{\mathbf{\Sigma}}$ to refine the predicted claim rates improves the overall fit of the model. Most importantly, it shows that moving from prior to (multivariate) posterior predicted claim rates

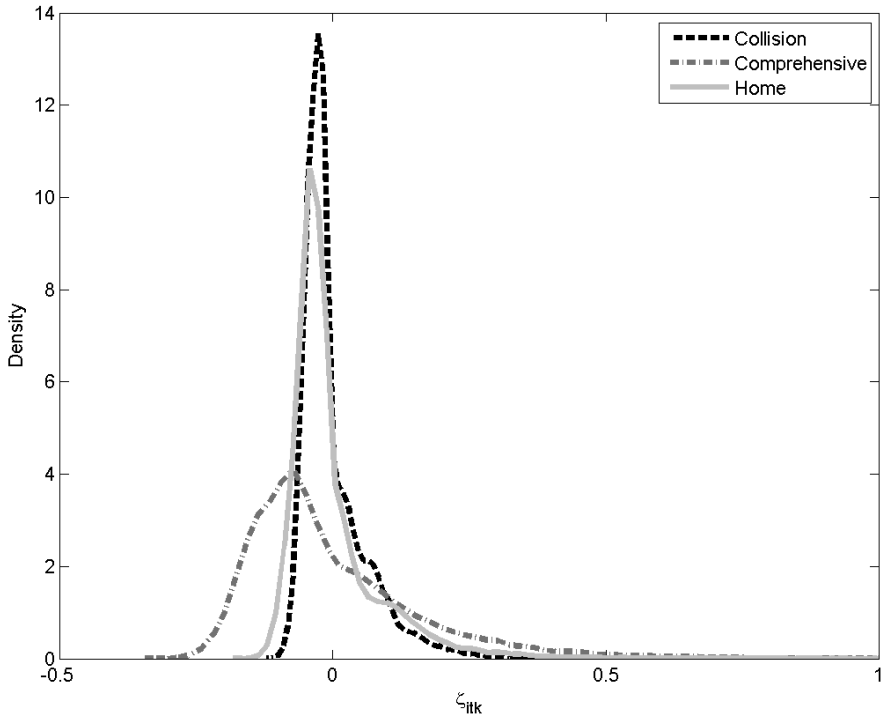

Figure 3.2: Kernel Density of ζ_{itk}

meliorates in each coverage the model's underprediction of zero counts and overprediction of one counts. In auto collision, the underprediction of zero counts decreases by 21 percent and the overprediction of one counts decreases by 13 percent. In auto comprehensive, the underprediction of zero counts decreases by 8 percent and the overprediction of one counts decreases by 5 percent. In home, the underprediction of zero counts decreases by 10 percent and the overprediction of one counts decreases by 6 percent.

Table 3.7: Descriptive Statistics for $\zeta = \frac{\theta - \vartheta}{\vartheta}$

	$\zeta < 0$				$\zeta > 0$			
Prior Claim Rates	Obs.	Mean	10^{th} Pctile	25^{th} Pctile	N	Mean	75^{th} Pctile	90^{th} Pctile
				Collision				
All	198,557	-0.034	-0.059	-0.046	96,360	0.069	0.093	0.153
Low	51,883	-0.030	-0.052	-0.041	21,846	0.067	0.092	0.148
Medium	99,078	-0.034	-0.059	-0.047	48,381	0.069	0.093	0.153
High	47,596	-0.038	-0.065	-0.052	26,133	0.070	0.095	0.157
				Comprehensive				
All	183,662	-0.098	-0.173	-0.139	111,255	0.157	0.217	0.357
Low	46,521	-0.095	-0.171	-0.135	27,208	0.163	0.224	0.370
Medium	91,981	-0.099	-0.175	-0.140	55,478	0.159	0.220	0.359
High	45,160	-0.098	-0.173	-0.139	28,569	0.149	0.207	0.334
				Home				
All	202,137	-0.044	-0.079	-0.059	92,780	0.093	0.132	0.211
Low	51,610	-0.040	-0.074	-0.055	22,119	0.093	0.135	0.213
Medium	101,001	-0.046	-0.082	-0.062	46,458	0.095	0.134	0.214
High	49,526	-0.044	-0.080	-0.059	24,203	0.090	0.126	0.203

3.5.2 Simulation Study

The previous section demonstrates that allowing for correlated random effects and utilizing the information about unobserved heterogeneity contained in the estimated variance-covariance matrix, $\widehat{\mathbf{\Sigma}}$, leads to material refinements of the predicted claim rates in the tricoverage sample and improves the model's overall fit to the data. In this section, we move from the actual data to simulated data. The principal advantage and virtue of the simulated data is that we can observe the "true" baseline claim rates, λ_{itk}, and "true" variance-covariance matrix, $\mathbf{\Sigma}$, neither of which is observable in the actual data. This allows us to examine more directly and more precisely the value of the information in $\mathbf{\Sigma}$,

Table 3.8: Distribution of Claim Counts - Actual versus Predicted with Posterior

Tricoverage Sample (294,917 household-years)

		Predicted		
Count	Actual	Prior	Posterior	Improvement
		Collision		
0	90.09	89.95	89.98	21.43
1	9.22	9.45	9.42	13.04
2	0.64	0.57	0.58	14.29
3	0.05	0.03	0.03	-
4	-	-	-	-
		Comprehensive		
0	96.95	96.83	96.84	8.33
1	2.88	3.09	3.08	4.76
2	0.16	0.07	0.07	-
3	0.01	-	-	-
4	-	-	-	-
		Home		
0	92.90	92.48	92.52	9.52
1	6.40	7.17	7.12	6.49
2	0.63	0.33	0.35	-
3	0.05	0.01	0.01	-
4	-	-	-	-

Note: Values are percentages. Predicted distributions are based on prior claim rates or multivariate posterior claim rates, as the case may be. Dot indicates less than 0.01 percent.

and in particular the independent value of the information on the cross-coverage correlation structure of unobserved heterogeneity, in terms of (i) improving the accuracy of the predicted claim rates and (ii) updating predicted claim rates to reflect subsequent claims experience.

Assumptions

We consider 18 cases. In each case, there N identical households. The cases differ on three variables: the households' baseline claim rates, $\lambda \equiv (\lambda_c, \lambda_m, \lambda_h)$; the time horizon, T; and the variance-covariance matrix, $\mathbf{\Sigma}$. We consider three levels of baseline claim rates: (1) "average" baseline claim rates, which correspond to the mean prior claim rates in the tricoverage sample: $\lambda = (0.100, 0.030, 0.070)$; (2) "low" baseline claim rates, which correspond to the 25th percentiles: $\lambda = (0.078, 0.016, 0.054)$; and (3) "high" baseline claim rates, which correspond to the 75th percentiles: $\lambda = (0.127, 0.044, 0.096)$. We also consider three time horizons: $T = 1; T = 3$; and $T = 10$. Finally, we consider two specifications for the variance-covariance matrix.[10] In specification A, we set $\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}$ to match the estimates from the data. In specification B, however, we harmonize the within-coverage variances such that each equals 0.250 and we adjust the cross-coverage covariances such that the cross-coverage correlations still match the estimates from the data (i.e., $\rho = \widehat{\rho}$). Hence, in specification A $\sigma^2 = (0.107, 0.399, 0.405)$ and $\rho = (0.663, 0.293, 0.559)$, which yields

$$\mathbf{\Sigma} = \begin{bmatrix} 0.107 & 0.137 & 0.061 \\ 0.137 & 0.399 & 0.225 \\ 0.061 & 0.225 & 0.405 \end{bmatrix} \text{ (specification A),}$$

and in specification B $\sigma^2 = (0.250, 0.250, 0.250)$ and $\rho = (0.663, 0.293, 0.559)$, which yields

$$\mathbf{\Sigma} = \begin{bmatrix} 0.250 & 0.166 & 0.073 \\ 0.166 & 0.250 & 0.140 \\ 0.073 & 0.140 & 0.250 \end{bmatrix} \text{ (specification B).}$$

The virtue of moving from specification A to specification B is that it changes the within-coverage variances (and harmonizes them at a "middle" level) while holding constant

[10] Of course, we fix $E([u_{ic}\ u_{im}\ u_{ih}]) = [1\ 1\ 1]$ in all cases.

the cross-coverage correlation structure. This serves to "identify" the independent value of the information in ρ, for we can determine the extent to which the results are—or are not—being driven by σ^2.

Accuracy

The first exercise demonstrates the independent value of the information in ρ in terms of improving the accuracy of the predicted claim rates. In this exercise, there are $10,000$ households and we perform 500 iterations of each case. In each iteration $j = 1, ..., 500$: (i) we simulate the (time-constant) random effect $\mathbf{u}_i$ for each household—i.e., we draw $\mathbf{u}_i \equiv [u_{ic}\ u_{im}\ u_{ih}]'$ from $lognormal([1\ 1\ 1]', \mathbf{\Sigma})$ independently for each household $i = 1, ..., 10,000$; (ii) we simulate the claims experience $\mathbf{y}_i$ of each household—i.e., for each household $i = 1, ..., 10,000$ and each year $t = 1, ..., T$, we draw y_{itk} from $Poisson\left(\lambda_k u_{ik}^j\right)$ for each coverage $k = c, m, l$); (iii) we estimate the model on the simulated data—i.e., we obtain $\widehat{\mathbf{\Sigma}}$; and (iv) and we calculate the univariate and multivariate posterior claim rates for each household—i.e., for each household $i = 1, ..., 10,000$, we calculate $\widehat{\vartheta}_{ik} \equiv \lambda_k E^{UV}(u_{ik}|\mathbf{y}_i)$ and $\widehat{\theta}_{ik} \equiv \lambda_k E^{MV}(u_{ik}|\mathbf{y}_i)$ for each coverage $k = c, m, l$.[11] We then compute, for each coverage $k = c, m, l$,

$$MSE_{ijk}^{UV} \equiv \frac{1}{10,000}\frac{1}{500}\sum_{i=1}^{10,000}\sum_{j=1}^{500}\left(\widehat{\vartheta}_{ijk} - \lambda_k u_{ijk}\right)^2$$

and

$$MSE_{ijk}^{MV} \equiv \frac{1}{10,000}\frac{1}{500}\sum_{i=1}^{10,000}\sum_{j=1}^{500}\left(\widehat{\theta}_{ijk} - \lambda_k u_{ijk}\right)^2,$$

[11] Strickly speaking, we should use $\bar{y}_{itk} \equiv \frac{1}{10,000}\frac{1}{T}\sum_{t=1}^{10,000}\sum_{t=1}^{T} y_{itk}$ instead of λ_k in calculating $\widehat{\vartheta}_{ik}$and $\widehat{\theta}_{ik}$, to make them directly comparable to $\widehat{\vartheta}_{itk}$ and $\widehat{\theta}_{itk}$. However, we use λ_k to abstract from the statistical uncertainty in estimating prior claim rates. The benefit is that we isolate the value of utilizing the information in $\widehat{\mathbf{\Sigma}}$. And the cost is small—by the law of large numbers, $\bar{y}_{itk} \approx \lambda_k$.

as well as

$$MSE_{ik}^{UV} \equiv \frac{1}{500}\sum_{j=1}^{500}\left(\widehat{\vartheta}_{ijk} - \lambda_k u_{ijk}\right)^2 \text{ and } MSE_{ik}^{MV} \equiv \frac{1}{500}\sum_{j=1}^{500}\left(\widehat{\theta}_{ijk} - \lambda_k u_{ijk}\right)^2$$

for each household $i = 1, ..., 10,000$ and

$$MSE_{jk}^{UV} \equiv \frac{1}{10,000}\sum_{i=1}^{10,000}\left(\widehat{\vartheta}_{ijk} - \lambda_k u_{ijk}\right)^2 \text{ and } MSE_{j}^{MV} \equiv \frac{1}{10,000}\sum_{i=1}^{10,000}\left(\widehat{\theta}_{ijk} - \lambda_k u_{ijk}\right)^2$$

for each iteration $j = 1, ..., 500$.

Table 3.9 summarizes the results. There are two main takeaways. The first main takeaway is that, in every case, $MSE_{ijk}^{MV} < MSE_{ijk}^{UV}$ for each coverage. In cases with a one-year time horizon, the multivariate posterior claim rates are only slightly more accurate than the univariate claim rates. In these cases, the absolute errors of the of the multivariate posterior claim rates are 0.1 percent to 1.0 percent less than the absolute errors of the univariate posterior claim rates. However, the accuracy advantage of the multivariate posterior claim rates is more pronounced in cases with higher baseline claim rates and longer time horizons. Indeed, in cases with high baseline claim rates and a ten-year time horizon, the absolute errors of the multivariate posterior claim rates are 2.0 percent to 7.6 percent less than the absolute errors of the univariate posterior claim rates. This increase in accuracy is due to the increase in the percentage of households for whom $MSE_{ik}^{MV} < MSE_{ik}^{UV}$, which increases from roughly 70 to 80 percent in cases with low claim rates and a one-year time horizon to more than 95 percent in cases with high claim rates and a ten-year time horizon. By contrast, $MSE_{jk}^{MV} < MSE_{jk}^{UV}$ in virtually every iteration of every case. The second main takeaway is that the accuracy advantage of the multivariate posterior claim rates is essentially the same for both specifications of $\boldsymbol{\Sigma}$. This suggests that, in terms of improving the accuracy of the predicted claim rates, the value of the information on the cross-coverage correlation structure of unobserved heterogeneity does not depend on the structure of the within-coverage variance of unobserved heterogeneity.

Table 3.9: Simulation Study - Accuracy

Time	Square Root of MSE_{ij}						% of Household where $MSE_i^{MV} < MSE_i^{UV}$			% of Household where $MSE_j^{MV} < MSE_j^{UV}$		
	Univariate			Multivariate								
Horizon	Home	Coll	Comp	Home	Coll	Comp	Home	Coll	Comp	Home	Coll	Comp
Σ Specification A												
Average Baseline Claim Rate:												
1	0.0439	0.0325	0.0188	0.0438	0.0324	0.0187	71.65	79.53	85.52	99.80	100.00	100.00
5	0.0416	0.0319	0.0184	0.0412	0.0313	0.0179	88.60	96.72	98.67	100.00	100.00	100.00
10	0.0392	0.0311	0.0179	0.0385	0.0302	0.0170	94.54	99.30	99.88	100.00	100.00	100.00
Low Baseline Claim Rate:												
1	0.0340	0.0254	0.0101	0.0340	0.0253	0.0100	66.79	74.18	82.57	97.80	100.00	100.00
5	0.0326	0.0250	0.0099	0.0324	0.0247	0.0097	83.71	93.06	97.70	100.00	100.00	100.00
10	0.0310	0.0245	0.0098	0.0307	0.0240	0.0094	90.01	97.42	99.63	100.00	100.00	100.00
High Baseline Claim Rate:												
1	0.0598	0.0412	0.0275	0.0596	0.0410	0.0273	76.01	84.48	89.06	99.80	100.00	100.00
5	0.0557	0.0402	0.0266	0.0550	0.0393	0.0256	91.90	98.44	99.36	100.00	100.00	100.00
10	0.0515	0.0389	0.0256	0.0505	0.0375	0.0240	96.18	99.71	99.90	100.00	100.00	100.00
Σ Specification B												
Average Baseline Claim Rate:												
1	0.0347	0.0494	0.0149	0.0346	0.0493	0.0148	73.87	73.38	89.76	100.00	99.80	100.00
5	0.0335	0.0471	0.0147	0.0332	0.0466	0.0142	90.86	91.25	99.63	100.00	100.00	100.00
10	0.0323	0.0447	0.0145	0.0317	0.0439	0.0136	95.81	96.27	99.98	100.00	100.00	100.00
Low Baseline Claim Rate:												
1	0.0268	0.0386	0.0080	0.0268	0.0385	0.0079	69.88	68.85	86.63	99.40	99.40	100.00
5	0.0261	0.0372	0.0079	0.0260	0.0370	0.0077	86.81	86.27	98.97	100.00	100.00	100.00
10	0.0253	0.0356	0.0078	0.0250	0.0352	0.0074	92.40	92.06	99.93	100.00	100.00	100.00
High Baseline Claim Rate:												
1	0.0473	0.0625	0.0219	0.0472	0.0623	0.0216	76.86	76.87	91.89	100.00	100.00	100.00
5	0.0453	0.0590	0.0214	0.0447	0.0581	0.0204	93.07	94.10	99.87	100.00	100.00	100.00
10	0.0430	0.0551	0.0209	0.0422	0.0538	0.0193	97.19	98.03	99.99	100.00	100.00	100.00

Updating

The second exercise explores the independent value of the information in ρ in terms of updating the predicted claim rates to reflect subsequent claims experience. We consider the same 18 cases, but in this exercise there is only one household and instead of generating 500 claim histories we consider four special histories: (i) the household experiences zero claims; (ii) the household experiences one claim in auto collision; (iii) the household experiences one claim in auto comprehensive; and (iv) the household experiences one claim in home. For each case, we then compute and compare the household's univariate and multivariate posterior claim rates for each history.

Table 3.10 presents the results. They display two key advantages of the multivariate approach relative to the univariate approach in terms of updating the household's predicted claim rates when it experiences a claim in one line of coverage. First and foremost, the multivariate approach updates the household's predicted claim rates both within and across coverages—i.e., the household's predicted claim rate in each coverage increases under the multivariate approach, whereas only its predicted claim rate for the coverage in which it experiences a claim increases under the univariate approach. Moreover, the magnitude of the cross-coverage updates are material, ranging from 4 percent to 22 percent. This highlights the fact that the univariate approach ignores valuable information—namely, the information on the cross-coverage correlation structure of unobserved heterogeneity contained in ρ. Second, the univariate approach materially "over-updates" the predicted claim rate for the coverage in which the household experiences a claim when the variance of unobserved heterogeneity within such coverage is low and the baseline claim rates are average or high. To see this, compare (1) the results for auto collision under specification A when the baseline claim rates are average and high with (2) the

results for auto collision under specification A when the baseline claim rates are low, (3) the results for auto comprehensive and home under specification A, and (4) the results for all three coverages under specification B. In the cases described in (1), the percentage increase in the unilateral posterior claim rate for auto collision in response to an auto collision claim is more than three times the percentage increase in the multivariate posterior claim rate for auto collision in response to an auto collision claim. In the other cases, however, the percentage increase in the unilateral posterior claim rate for auto collision in response to an auto collision claim is roughly equal to the percentage increase in the multivariate posterior claim rate for auto collision in response to an auto collision claim. This suggests that, in terms of updating the household's predicted claim rates to reflect subsequent claims experience, the value of the information on the cross-coverage correlation structure of unobserved heterogeneity depends on the within-coverage variance of unobserved heterogeneity (as well as the baseline claim risk), and is most valuable when this variance is low (and when baseline claim risk is average or high).

Table 3.10: Simulation Study - Updating

Time Horizon	Baseline Claim Rate	Method	Coverage	Prior Claim Rate	Posterior claim rates with: Zero Claims	Posterior claim rates with: One claim in: Home	Coll	Comp	Posterior claim rate % increase due to one claim in: Home	Coll	Comp
Σ Specification A											
1	Avg.	MV	Home	0.070	0.067	0.093	0.071	0.082	38.78	5.79	21.40
			Coll	0.100	0.098	0.104	0.108	0.111	5.70	10.39	13.24
			Comp	0.030	0.029	0.035	0.033	0.040	21.53	13.19	38.89
		UV	Home	0.070	0.068	0.095	.	.	39.06	.	.
			Coll	0.100	0.096	.	0.133	.	.	38.46	.
			Comp	0.030	0.030	.	.	0.042	.	.	40.20
	Low	MV	Home	0.054	0.052	0.073	0.056	0.064	39.31	5.92	21.76
			Coll	0.078	0.077	0.082	0.085	0.087	5.84	10.52	13.38
			Comp	0.016	0.016	0.019	0.018	0.022	21.15	12.82	38.46
		UV	Home	0.054	0.054	0.059	.	.	10.61	.	.
			Coll	0.078	0.077	.	0.086	.	.	10.47	.
			Comp	0.016	0.016	.	.	0.018	.	.	10.63
	High	MV	Home	0.096	0.091	0.126	0.096	0.110	38.24	5.71	20.99
			Coll	0.127	0.124	0.131	0.137	0.140	5.65	10.41	13.16
			Comp	0.044	0.042	0.051	0.047	0.058	21.10	13.19	38.13
		UV	Home	0.096	0.093	0.128	.	.	38.05	.	.
			Coll	0.127	0.121	.	0.166	.	.	37.41	.
			Comp	0.044	0.043	.	.	0.060	.	.	39.12
5	Avg.	MV	Home	0.070	0.059	0.079	0.062	0.070	34.28	4.89	18.40
			Coll	0.100	0.092	0.096	0.101	0.103	4.87	9.83	11.98
			Comp	0.030	0.025	0.030	0.028	0.034	18.38	12.01	34.92
		UV	Home	0.070	0.062	0.083	.	.	34.75	.	.
			Coll	0.100	0.085	.	0.113	.	.	33.15	.
			Comp	0.030	0.028	.	.	0.039	.	.	37.61
	Low	MV	Home	0.054	0.047	0.064	0.050	0.056	35.46	5.16	19.32
			Coll	0.078	0.073	0.077	0.081	0.082	5.16	10.05	12.45
			Comp	0.016	0.014	0.017	0.016	0.019	19.37	12.54	36.47
		UV	Home	0.054	0.053	0.058	.	.	10.36	.	.
			Coll	0.078	0.075	.	0.083	.	.	10.25	.
			Comp	0.016	0.016	.	.	0.018	.	.	10.59
	High	MV	Home	0.096	0.077	0.102	0.080	0.090	32.81	4.56	17.41
			Coll	0.127	0.114	0.119	0.125	0.127	4.58	9.63	11.56
			Comp	0.044	0.035	0.041	0.039	0.047	17.39	11.56	33.70
		UV	Home	0.096	0.082	0.109	.	.	32.96	.	.
			Coll	0.127	0.104	.	0.137	.	.	31.56	.
			Comp	0.044	0.041	.	.	0.055	.	.	35.99

Continued on next page...

Time Horizon	Baseline Claim Rate	Method	Coverage	Prior Claim Rate	Posterior claim rates with: Zero Claims	Posterior claim rates with: One claim in: Home	Coll	Comp	Posterior claim rate % increase due to one claim in: Home	Coll	Comp
Σ Specification A											
10	Average	MV	Home	0.070	0.052	0.068	0.054	0.060	30.96	4.17	16.21
			Coll	0.100	0.086	0.089	0.094	0.095	4.18	9.27	10.94
			Comp	0.030	0.022	0.025	0.024	0.029	16.20	10.95	32.13
		UV	Home	0.070	0.056	0.074	.	.	31.42	.	.
			Coll	0.100	0.075	.	0.097	.	.	29.40	.
			Comp	0.030	0.027	.	.	0.036	.	.	35.40
	Low	MV	Home	0.054	0.043	0.056	0.045	0.050	32.42	4.55	17.39
			Coll	0.078	0.069	0.073	0.076	0.077	4.57	9.57	11.58
			Comp	0.016	0.013	0.015	0.014	0.017	17.43	11.57	34.07
		UV	Home	0.054	0.051	0.056	.	.	10.07	.	.
			Coll	0.078	0.072	.	0.079	.	.	9.84	.
			Comp	0.016	0.016	.	.	0.017	.	.	10.49
	High	MV	Home	0.096	0.066	0.085	0.068	0.075	29.24	3.83	15.05
			Coll	0.127	0.105	0.109	0.114	0.116	3.82	8.97	10.38
			Comp	0.044	0.030	0.034	0.033	0.039	15.03	10.39	30.56
		UV	Home	0.096	0.073	0.094	.	.	29.32	.	.
			Coll	0.127	0.090	.	0.115	.	.	27.68	.
			Comp	0.044	0.038	.	.	0.051	.	.	33.36
Σ Specification B											
1	Avg.	MV	Home	0.070	0.068	0.085	0.073	0.077	24.23	6.90	13.36
			Coll	0.100	0.097	0.103	0.120	0.112	6.93	24.10	15.93
			Comp	0.030	0.029	0.033	0.034	0.036	13.45	16.21	24.48
		UV	Home	0.070	0.069	0.086	.	.	24.56	.	.
			Coll	0.100	0.098	.	0.121	.	.	24.28	.
			Comp	0.030	0.030	.	.	0.037	.	.	24.83
	Low	MV	Home	0.054	0.053	0.066	0.057	0.060	24.39	6.99	13.61
			Coll	0.078	0.076	0.081	0.095	0.088	7.11	24.47	16.18
			Comp	0.016	0.016	0.018	0.018	0.020	13.46	16.03	25.00
		UV	Home	0.054	0.053	0.066	.	.	24.58	.	.
			Coll	0.078	0.077	.	0.095	.	.	24.44	.
			Comp	0.016	0.016	.	.	0.020	.	.	25.16
	High	MV	Home	0.096	0.092	0.115	0.099	0.105	24.24	6.93	13.31
			Coll	0.127	0.122	0.130	0.151	0.141	6.83	23.95	15.80
			Comp	0.044	0.042	0.048	0.049	0.052	13.30	15.91	24.23
		UV	Home	0.096	0.094	0.117	.	.	24.31	.	.
			Coll	0.127	0.123	.	0.153	.	.	24.03	.
			Comp	0.044	0.044	.	.	0.054	.	.	24.83

Continued on next page...

Time Horizon	Baseline Claim Rate	Method	Coverage	Prior Claim Rate	Posterior claim rates with: Zero Claims	Posterior claim rates with: One claim in: Home	Coll	Comp	Posterior claim rate % increase due to one claim in: Home	Coll	Comp
					Σ Specification B						
5	Avg.	MV	Home	0.070	0.062	0.076	0.065	0.069	22.42	5.86	12.02
			Coll	0.100	0.086	0.091	0.105	0.098	5.86	21.68	14.03
			Comp	0.030	0.026	0.029	0.030	0.032	12.03	14.03	22.28
		UV	Home	0.070	0.065	0.079	.	.	22.75	.	.
			Coll	0.100	0.090	.	0.109	.	.	22.02	.
			Comp	0.030	0.029	.	.	0.036	.	.	23.91
	Low	MV	Home	0.054	0.049	0.060	0.052	0.055	22.91	6.16	12.43
			Coll	0.078	0.069	0.074	0.085	0.080	6.19	22.34	14.59
			Comp	0.016	0.014	0.016	0.016	0.018	12.55	14.64	23.01
		UV	Home	0.054	0.051	0.062	.	.	23.23	.	.
			Coll	0.078	0.071	.	0.088	.	.	22.58	.
			Comp	0.016	0.016	.	.	0.020	.	.	24.49
	High	MV	Home	0.096	0.081	0.099	0.086	0.091	21.75	5.55	11.51
			Coll	0.127	0.105	0.111	0.127	0.119	5.53	21.02	13.48
			Comp	0.044	0.036	0.041	0.041	0.044	11.51	13.49	21.60
		UV	Home	0.096	0.086	0.105	.	.	22.11	.	.
			Coll	0.127	0.111	.	0.135	.	.	21.44	.
			Comp	0.044	0.042	.	.	0.052	.	.	23.46
10	Average	MV	Home	0.070	0.056	0.068	0.059	0.062	20.78	5.03	10.82
			Coll	0.100	0.077	0.081	0.092	0.086	5.02	19.75	12.56
			Comp	0.030	0.023	0.026	0.026	0.028	10.84	12.56	20.68
		UV	Home	0.070	0.060	0.073	.	.	21.18	.	.
			Coll	0.100	0.082	.	0.098	.	.	20.14	.
			Comp	0.030	0.028	.	.	0.034	.	.	23.04
	Low	MV	Home	0.054	0.045	0.055	0.048	0.050	21.52	5.43	11.43
			Coll	0.078	0.063	0.067	0.076	0.072	5.44	20.58	13.30
			Comp	0.016	0.013	0.015	0.015	0.016	11.43	13.34	21.65
		UV	Home	0.054	0.048	0.058	.	.	21.85	.	.
			Coll	0.078	0.066	.	0.080	.	.	20.90	.
			Comp	0.016	0.015	.	.	0.019	.	.	23.83
	High	MV	Home	0.096	0.072	0.086	0.075	0.079	19.88	4.60	10.17
			Coll	0.127	0.092	0.096	0.109	0.102	4.62	18.93	11.87
			Comp	0.044	0.032	0.035	0.036	0.038	10.20	11.87	19.83
		UV	Home	0.096	0.079	0.095	.	.	20.27	.	.
			Coll	0.127	0.100	.	0.119	.	.	19.36	.
			Comp	0.044	0.040	.	.	0.049	.	.	22.30

3.6 Conclusion

There is significant statistical and economic value in utilizing within-coverage variance and cross-coverage correlation of unobserved heterogeneity in modeling and predicting claim risk. Empirical analysis of the insurance data shows that the information contained in $\widehat{\mathbf{\Sigma}}$ leads to material refinements of predicted claim rates; and simulation studies establish the independent value of cross-coverage information in improving the accuracy of predicted claim rates and usefulness in updating predicted claim rates to reflect claim experience. In addition to illustrating desirable statistical and economic advantages, these findings suggest that legal restrictions on experience rating exacerbate any dead weight loss from adverse selection. Future work will estimate dead weight loss using techniques developed by Einav et al. (2010).

CHAPTER 4

VARIATIONAL APPROXIMATE INFERENCE FOR JOINT MODELING OF MULTIVARIATE LONGITUDINAL AND DURATION DATA

4.1 Introduction

Joint correlated longitudinal and duration data arise when multiple outcomes are measured repeatedly on a subject over time along with time-to-event outcomes. This type of data is common in longitudinal studies in biostatistics and observational panel data in economics. For example, joint longitudinal and duration data naturally arise in the unbalanced panel setting, where a set of repeated measurements is of primary interest, but time-to-dropout from the sample may be treated as a secondary outcome. Separate analysis of the longitudinal and duration outcomes is straightforward and well-established in the literature, but it is useful to develop a framework and estimation techniques for modeling the disparate outcomes together. Joint modeling is particularly important when the research question of interest concerns the dependence between the longitudinal and duration outcomes.

Research in joint modeling of longitudinal and duration data grew out of interest in dealing with informative dropouts: their connection to the hierarchy of missing data mechanisms set forth by Little and Rubin (2002) and the converse problem of modeling a time-to-event outcome with mis-measured covariates originally studied by Tsiatis et al. (1995). A joint modeling framework is necessary for addressing research questions that focus in varying degrees on the longitudinal model and/or the duration model. The researcher may be interested in (1) characterizing the relationship between the longitudinal process and the duration outcome, (2) accounting for complications of dropout in

longitudinal outcomes, and/or (3) addressing the effect of time-varying covariates in a duration model. Our research is motivated by interest in the first two objectives. In particular, we want to develop a model and methodology to assess the potential for attrition bias by characterizing the role that unobserved time-constant subject-specific characteristics play in a subject's propensity to be observed in the sample. We are also interested in the role that time-to-dropout plays in estimating unobserved subject-specific effects.

Various frameworks have been proposed for joint models of disparate outcomes that characterize the systematic relationships in different ways. Correlated random effects models have received a lot of attention in the literature due to their flexibility and extendibility to higher dimensions. This research adopts a version of the random effects model proposed by Wulfsohn and Tsiatis (1997), in which the longitudinal measurements and duration outcomes are assumed conditionally independent given a set of subject-specific unobserved random effects. In this model, the correlated random effects induce dependence between the longitudinal and duration outcomes. While the literature has focused on joint models of univariate longitudinal and univariate duration outcomes, correlated random effects models can naturally be extended to jointly consider multivariate longitudinal outcomes and/or multivariate duration outcomes.

In theory, correlated random effects models are not limited by the dimension of the random effects vector; but in practice there are computational limitations and challenges with maximum likelihood estimation since the marginal likelihood involves an intractable integral. In fact, the multivariate longitudinal submodel alone poses computational challenges that warrant the development of estimation techniques (Fitzmaurice et al., 2009; Fieuws and Verbeke, 2006; Morris, 2011). A number of estimation techniques have been proposed to overcome the computational complexity of direct maximum likelihood estimation in joint random effects models, including: naive two-stage,

EM algorithm (Wulfsohn and Tsiatis, 1997), conditional score (Tsiatis and Davidian, 2004), Laplace approximation (Rizopulos et al., 2009), Markov chain Monte Carlo (MCMC), and Bayesian methods. A good review of these approaches, including advantages and disadvantages of each, can be found in Wu et al. (2012). Previous research has focused on univariate cases of linear mixed models and proportional hazard survival models. Many of these proposed methods become even more computationally challenging when extending to multivariate, i.e. multiple longitudinal outcomes and/or multiple duration outcomes, and non-normal longitudinal outcomes, i.e. incorporating a generalized linear mixed model for the measurement submodel.

We propose using Gaussian variational approximation (GVA) to overcome the computational complexity of a multivariate longitudinal count and multivariate duration random effects model. Variational approximation, as summarized in Jordan et al. (1999), Jordan (2004), and Bishop (2006), is an established, contemporary methodology in computer science and machine learning. Ormerod and Wand (2010) introduce variational approximation to the statistical modeling literature as a fast and deterministic alternative to MCMC methods for integration, which sacrifices accuracy for computational feasibility. While variational approximation has been largely applied to Bayesian models, GVA has recently been shown to perform well as an approximate estimation technique in the grouped generalized linear mixed models context (Ormerod and Wand, 2011). We extend this methodology to the case of joint longitudinal count and duration models, where high-dimensional intractable integrals are of concern. The random effects model proposed in this research relies on distributional assumptions associated with the longitudinal outcomes, duration outcomes and random effects marginal distribution. Additionally, the GVA method imposes assumptions on an approximate posterior distribution of the random effects conditional on the data. The validity of such assumptions may be difficult to

assess on the observable data, but lead to a less computationally demanding estimation technique.

This research contributes to the literature in joint modeling by providing a novel extension of the joint modeling framework to multivariate and non-normal data, and introducing GVA as a computationally advantageous estimation technique for joint models. GVA is comparable in computational complexity to naive two-stage models, offering significant computational advantages over numerical methods. With respect to estimation of association parameters in the joint model, in most cases, we find that both approaches perform better than separate modeling, with GVA exhibiting better properties than a multivariate two-stage approach when correlation of unobserved heterogeneity is not present. In addition, the GVA approach is a fully joint model, as opposed to the two-stage approach, in that the measurement and duration submodels are estimated together in one stage. This property allows estimation of quantities, e.g. posterior expectation of unobserved heterogeneity that is conditioned on both the count and the duration outcomes, that the two-stage approach does not and would be computationally prohibitive with numerical methods.

The rest of this chapter is organized as follows: Section 4.2 introduces the motivating research question and the insurance data; Section 4.3 describes the model and Gaussian variational approximation; Section 4.4 presents simulation studies demonstrating finite sample properties; Section 4.5 discusses the main empirical findings; Section 4.6 reviews important alternative joint models, identifiability of variational parameters, and computational considerations; and Section 4.7 concludes.

4.2 Motivating Example: Insurance Data

This research is motivated by an empirical question concerning the association between unobserved heterogeneity in multiple insurance claim count processes and policy duration. We are interested in whether the unobserved "riskiness" of a policyholder is associated with propensity to maintain a policy in force. If so, then ignoring this association in a multivariate longitudinal claim count model impacts properties of the estimator of covariate and association effects, as "riskier" policies may be more or less likely to be observed. We want to assess the association between "riskiness" and risk for dropout. Additionally, incorporating duration information can lead to improvements in estimation of unobserved heterogeneity. Random effects in a longitudinal claim count model for multiple coverages serve as a measure of the inherent time-constant "riskiness" of a policyholder which we would like to relate to policy duration. This research question extends that of Chapter 2 (Morris, 2011) and Chapter 3 (Barseghyan et al., 2012) to address concerns about attrition bias.

The motivating dataset, acquired from a large U.S. property and casualty insurance company, contains yearly information for multiple lines of personal insurance coverage. This dataset contains household level matched records for home and auto insurance observed over the course of nine years, 1998 - 2006. At the beginning of each year, we observe a snapshot of policy and household characteristics, such as insurance score, that are linked to the number of claims filed during the course of the year. The unbalanced panel sample of $27,051$ policies includes those households that have a complete set of claim count outcomes and covariates for all three coverages at any point in the nine year period, i.e. both home and auto policies in force in any year from 1998-2006. It also only includes those policies originated during the period of observation. This results in a total

of 98,804 policy/year observations. The count outcomes of interest are:

Longitudinal Measurement Outcomes:

$$y_{it1} = \text{number of home claims for policy } i \text{ and time period } t$$
$$y_{it2} = \text{number of collision claims for policy } i \text{ and time period } t$$
$$y_{it3} = \text{number of comprehensive claims for policy } i \text{ and time period } t$$

The data also includes origination and cancellation dates for both home and auto insurance policies for a given customer. The duration analysis dependent variables of interest are:

Duration Outcomes:

$$T_{i1} = \text{observed time to cancellation of home policy for policy } i$$
$$T_{i2} = \text{observed time to cancellation of auto policy for policy } i$$
$$\delta_{i1} = \text{indicator of observed home policy cancellation for policy } i$$
$$\delta_{i2} = \text{indicator of observed auto policy cancellation for policy } i$$

Cancellation is observed for about 25% of the policies and the average duration is about 5.15 and 5.25 years for home and auto policies, respectively (see Table 4.1). Policies are assumed to remain out of force once canceled and censoring only occurs at the end of the observation period, i.e. all policies are censored in 2006. While the longitudinal measurement submodel includes only data observed when both home and auto coverages are in force, policy duration is calculated for the full observation period for each policy type separately. Cancellation is defined as either voluntary or involuntary.[1]

[1]While we have information on cancellation reason in the insurance data, we have not incorporated this

Table 4.1: Summary Statistics for Insurance Claim Counts and Policy Duration

				Mean # of Claims per Year		
Insurance Type	Mean Duration	% Censored	% Attrition	Overall	Censored	Canceled
Home	5.15 years	76.6	23.4	.079	.076	.087
Collision	5.25 years	74.1	25.9	.108	.095	.150
Comprehensive				.032	.031	.037

Note: Includes 98,804 policy/year and 27,051 policy observations.
Mean number of claims taken over all years, but vary only slightly by year.

It is important to note that there is a naturally strong dependence between the two duration outcomes: of the policies for which we observed a cancellation, about 45% cancel both home and auto, about 31% cancel only their auto policy, and about 24% cancel only their home policy during the period of observation. About 26% cancel both home and auto coverages at the same time.

Separate modeling of both the longitudinal measurement and the duration components is straightforward, but our interest is in the association between unobserved heterogeneity in the count model and the policy duration. The relation between the policy duration outcomes and the claim count outcomes through shared random effects is a particularly important research question since the method for multivariate longitudinal count submodel alone assumes no association between the processes, i.e. attrition is random. Table 4.1 provides preliminary evidence of association between claim rates and policy duration through differences in unconditional annual claim rates between the overall sample and the sample of policies that cancel one or both policy types. A joint model addresses questions of the dependence between claim experience in three types of

information in this research. The distinction between voluntary/involuntary cancellation, i.e. cancellation at insured/company's request, will be explored in future research.

coverage and propensity to maintain two types of policies with the company.

4.3 Methodology

4.3.1 Notation and Model

For the measurement submodel, let y_{itk} denote the k^{th} count outcome and $\mathbf{x}_{itk}$ denote the p_kx1 vector of covariates observed for the k^{th} count and the i^{th} subject in time period t, where $i = 1, \ldots, N$, $t = 1, \ldots, T_i$ and $k = 1, \ldots, K$. Let $\mathbf{y}_{ik}$, λ_{ik}, and $\mathbf{x}_{ik}$ denote the T_ix1 vectors and T_ixp_k matrix of all measurements for the k^{th} outcome for the i^{th} subject, e.g. $\mathbf{y}_{ik} = \begin{bmatrix} y_{i1k} & \ldots & y_{iT_ik} \end{bmatrix}^T$. Let $\mathbf{y}_i$, λ_i denote the KT_ix1 vectors of all measurements for the i^{th} subject, e.g. $\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{i1}^T & \ldots & \mathbf{y}_{iK}^T \end{bmatrix}^T$.

For the duration submodel, let T_{ij} denote the j^{th} observed duration outcome and $\mathbf{z}_{ij}$ denote the p_jx1 vector of covariates observed for the j^{th} duration and the i^{th} subject, where $j = 1, \ldots, J$. Let T_{ij}^* denote the j^{th} underlying true event time and C_{ij} denote the censoring time for the j^{th} duration and the i^{th} subject, so that $T_{ij} = min(T_{ij}^*, C_{ij})$ and $\delta_{ij} = I[T_{ij}^* < C_{ij}]$. Let stacked vector notation follow from that defined for the measurement submodel.

To extend the joint longitudinal and duration random effects model to the multivariate count and duration setting, assume $\mathbf{b}_i$ to be the vector of correlated subject-specific latent effects for subject i with elements $(b_{i1}, \ldots, b_{iK})$ and the following regression equations and conditional distributions for the measurement and the duration submodels:

Measurement Submodel and Distributional Assumptions:

$$E(y_{itk}|\mathbf{x}_{itk}, b_{ik}) = e^{b_{ik}}\lambda_{ik}$$

$$\mathbf{y}_i|\mathbf{x}_i, \mathbf{b}_i = \begin{pmatrix} \mathbf{y}_{i1}|\mathbf{x}_{i1}, b_{i1} \\ \vdots \\ \mathbf{y}_{iK}|\mathbf{x}_{iK}, b_{iK} \end{pmatrix} \sim Poisson \begin{pmatrix} e^{b_{i1}}\lambda_{i1} \\ \vdots \\ e^{b_{iK}}\lambda_{iK} \end{pmatrix}$$

Duration Submodel and Distributional Assumptions:

$$h(t|\mathbf{z}_{ij}, \mathbf{b}_i) = h_0(t)e^{\mathbf{b}_i^T \alpha_j}\xi_{ij}$$

$$\mathbf{T}_i^*|\mathbf{z}_i, \mathbf{b}_i = \begin{pmatrix} T_{i1}^*|z_{i1}, \mathbf{b}_i \\ \vdots \\ T_{iJ}^*|z_{iJ}, \mathbf{b}_i \end{pmatrix} \sim Weibull \begin{pmatrix} r_1,\ e^{\mathbf{b}_i^T \alpha_1}\xi_{i1} \\ \vdots \\ r_J,\ e^{\mathbf{b}_i^T \alpha_J}\xi_{iJ} \end{pmatrix}$$

where h_0 is the form of the baseline hazard associated with the Weibull model[2], i.e. $h_0(t_{ij}) = r_j(t_{ij})^{r_j - 1}$, $\lambda_{itk} = exp\left(\mathbf{x}_{itk}^T \beta_k\right)$ and $\xi_{ij} = exp\left(\mathbf{z}_{ij}^T \gamma_j\right)$. That is, $\mathbf{y}_i$ follows a Poisson distribution conditional on a set of random effects, a set of covariates and a vector of regression parameters $(\beta_1, ..., \beta_K)$, which includes an intercept, that are common to all subjects; and $\mathbf{T}_i^*$ follows a Weibull distribution conditional on a set of random effects, a set of covariates and a vector of regression parameters $(\gamma_1, ..., \gamma_J)$, which includes an intercept, that are common to all subjects. The measurement and duration models are assumed conditionally independent given the unobserved time-constant random effect vector $\mathbf{b}_i$.[3]

Assuming conditional independence and accounting for censoring of the true underlying time-to-event outcome, the conditional densities can be written as:

[2]The hazard function is of primary interest in econometric duration models (Heckman and Leemer, 2001). Proportional hazard models are predominant in the econometrics literature because the hazard function is the focus of such models, as opposed to accelerated failure time (AFT) models. While the Weibull model for duration outcomes can be formulated equivalently as proportional hazards and AFT models, this research uses the proportional hazards specification.

[3]One or more components of the random effect vector from the measurement submodel need not be linked to the duration submodel, i.e. $\mathbf{b}_i$ in the duration submodel can be a subset of $\mathbf{b}_i$ in the measurement model. In this research we maintain all random effects terms in the duration submodel, as our research question dictates interest in all of these effects.

Measurement Submodel Density:

$$f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{b}_i) = \prod_{k=1}^{K} \prod_{t=1}^{T_i} \frac{e^{-e^{x_{itk}^T \beta_k + b_{ik}}} e^{y_{itk}(x_{itk}^T \beta_k + b_{ik})}}{y_{itk}!}$$

Duration Submodel Density:

$$f\left((\mathbf{T}_i, \delta_i)|\mathbf{z}_i, \mathbf{b}_i\right) = \prod_{j=1}^{J} \left(\left(r_j T_{ij}^{r_j - 1} e^{\mathbf{b}_i^T \alpha_j + z_{ij}^T \gamma_j} \right)^{\delta_{ij}} e^{-T_{ij}^{r_j} e^{\mathbf{b}_i^T \alpha_j + z_{ij}^T \gamma_j}} \right)$$

Furthermore, in this research we make a common distributional assumption that $\mathbf{b}_i$ follows a K-dimensional multivariate normal distribution with mean zero and covariance matrix Σ. That is:

$$f(\mathbf{b}_i|\Sigma) \quad = \quad (2\pi)^{-K/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}\mathbf{b}_i^T \Sigma^{-1} \mathbf{b}_i}$$

Taken jointly, the marginal density of $\mathbf{y}_i$ and $(\mathbf{T}_i, \delta_i)$ can be written as[4]:

$$L_i \quad = \quad \int_{b_{iK}} ... \int_{b_{i1}} f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{b}_i)\ f\left((\mathbf{T}_i, \delta_i)|\mathbf{z}_i, \mathbf{b}_i\right)\ f(\mathbf{b}_i|\Sigma)\ db_{i1}...db_{iK} \tag{4.1}$$

where maximum likelihood estimators of the parameters of interest, $(\beta, \Sigma, \alpha, \gamma)$, are defined as:

$$(\hat{\beta}, \hat{\Sigma}, \hat{\gamma}, \hat{\alpha}) = \underset{(\beta, \Sigma, \gamma, \alpha)}{\operatorname{argmax}} \prod_{i=1}^{N} L_i(\beta, \Sigma, \gamma, \alpha)$$

The marginal likelihood L_i in Equation 4.1 involves a possibly high-dimensional intractable integral. Numerical integration and direct maximization of this marginal likelihood can be computationally prohibitive, particularly as the dimension of the data increases. Various techniques have been proposed for estimation of such integrals encountered in joint modeling, including: Bayesian MCMC (Guo and Carlin, 2004), EM algorithm (Wulfsohn and Tsiatis, 1997) and conditional score (Tsiatis and Davidian, 2004).

[4]Under the specification that the random effects coefficients in the duration submodel, α, are equal to zero, all dependence between the measurement submodel and the duration submodel is eliminated and the joint model factors into the product of a grouped generalized linear mixed model and standard Weibull proportional hazards model.

These proposed methods have focused on the linear mixed model as the measurement submodel. For the generalized linear mixed model case, many of these methods present additional challenges. For example, the EM algorithm will not have a closed form in the expectation step. In this research we extend Gaussian variational approximation techniques (Ormerod and Wand, 2010, 2011) to estimate the joint multivariate longitudinal and duration random effects model.

4.3.2 Variational Approximation

Gaussian variational approximation (GVA) is a technique that avoids multivariate intractable integrals through assumptions on the posterior distributions of the random effects. Generally, GVA involves obtaining a variational lower bound for the log-likelihood of interest, by introducing variational parameters associated with an approximation of the posterior distribution. The variational estimator is then obtained by maximizing the variational lower bound. In the case of the joint count and duration model described in Section 4.3.1, GVA results in a closed form for the variational lower bound. The general variational lower bound for a joint random effects model for longitudinal and duration outcomes can be derived as:

$$\begin{aligned}
l(\beta, \Sigma, \gamma, \alpha) &= \log\left(\int f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)db\right) \\
&= \log\left(\int f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)\frac{q(b|\omega)}{q(b|\omega)}db\right) \\
&= \log E_q\left(\frac{f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)}{q(b|\omega)}\right) \\
&\geq E_q\left(\log\frac{f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)}{q(b|\omega)}\right) \\
&= l^*(\beta, \Sigma, \gamma, \alpha, \omega)
\end{aligned}$$

where $q(b|\omega)$ is the assumed posterior distribution of the random effects, ω is a set of variational parameters and inequality is introduced by Jensen's inequality. This inequality can also be derived through the notion of Kullback-Leibler distance between $q(b|\omega)$ and $f(b|y, T, \delta)$, so that the maximization of the variational lower bound coincides with the minimization of the Kullback-Leibler distance between the assumed approximate posterior distribution of the random effects and the true posterior distribution of the random effect. This can be seen through the following general derivation:

$$
\begin{aligned}
l(\beta, \Sigma, \gamma, \alpha) &= \log\left(\int f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)db\right)\left[\int q(b|\omega)db\right] \\
&= \int q(b|\omega)\left[\log\left(\int f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)db\right)\right]db \\
&= \int q(b|\omega)\left[\log\left(\frac{f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)}{f(b|y,T,\delta)} \times \frac{q(b|\omega)}{q(b|\omega)}\right)\right]db \\
&= \int q(b|\omega)\left[\log\left(\frac{f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)}{q(b|\omega)}\right)\right]db + \int q(b|\omega)\left[\log\left(\frac{q(b|\omega)}{f(b|y,T,\delta)}\right)\right]db \\
&\geq \int q(b|\omega)\left[\log\left(\frac{f(y|b,\beta)\ f((T,\delta)|b,\alpha,\gamma)\ f(b|\Sigma)}{q(b|\omega)}\right)\right]db \\
&= l^*(\beta, \Sigma, \gamma, \alpha, \omega)
\end{aligned}
$$

where inequality arises since the Kullback-Leibler divergence between $q(b|\omega)$ and $f(b|y, T, \delta)$ is non-negative (Kullback and Leibler, 1951). This derivation explicitly characterizes the dependence in the difference between the log-likelihood and the variational lower bound as the Kullback-Leibler divergence between the true posterior distribution of the random effects and the proposed approximate posterior.

4.3.3 Variational Lower Bounds for Joint Model

There are many important assumptions on which the variational lower bound for the joint model is based. As outlined in Section 4.3.1, we have assumed a Poisson distribution for the count outcomes conditional on the random effects, a Weibull distribution of the duration outcomes conditional on the random effects, and a multivariate normal distribution for the marginal distribution of the random effects. The essence of the GVA approach is assuming the approximation of the posterior distribution of the random effects, $f(b|y, T, \delta)$, by a distribution, $q(b|\omega)$, for which the likelihood given this assumption is tractable. In the case of the joint model, tractability is achieved by restricting the posterior distribution to a more manageable class of distributions that relies on assumptions of (1) factorization and (2) normality. Specifically, we assume that the posterior distribution of the random effects can be approximated by:

$$q(\mathbf{b}_1, \ldots, \mathbf{b}_N|\mu, \mathbf{\Lambda}) = \prod_{i=1}^{N} \phi(\mathbf{b}_i|\mu_i, \Lambda_i)$$

where μ_i and Λ_i are the variational parameters introduced for each subject, $\phi(\mathbf{b}_i|\mu_i, \Lambda_i)$ is a multivariate normal density with Kx1 mean vector μ_i and KxK covariance matrix Λ_i. We have partitioned the random effects vector by subject and imposed a multivariate normal distribution for the posterior distribution of a subject's random effects vector. This specific partitioning assumption is reasonable when subjects are assumed independent.

With these assumptions, the variational lower bound for the multivariate longitudinal count and multivariate duration model has a closed form:

$$
\begin{aligned}
l^*(\beta, \Sigma, \gamma, \alpha, \mu, \Lambda) = & \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T_i} \left[y_{itk} \left(\mathbf{x}_{itk}^T \beta_k + \mu_{ik} \right) - log(y_{itk}!) - e^{\mathbf{x}_{itk}^T \beta_k + \mu_{ik} + \frac{1}{2} \Lambda_i^{(kk)}} \right] \\
& + \sum_{i=1}^{N} \sum_{j=1}^{J} \left[\delta_{ij} \left(\log(r_j) + r_j \log(T_{ij}) - \log(T_{ij}) \right) \right] \\
& + \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij} \left(\mathbf{z}_{ij}^T \gamma_j + \mu_i^T \alpha_j \right) - \sum_{i=1}^{N} \sum_{j=1}^{J} \left(T_{ij}^{r_j} e^{\mathbf{z}_{ij}^T \gamma_j + \mu_i^T \alpha_j + \frac{1}{2} \alpha_j^T \Lambda_i \alpha_j} \right) \\
& + \frac{1}{2} \sum_{i=1}^{N} \log |\Sigma^{-1} \Lambda_i| - \frac{1}{2} \sum_{i=1}^{N} \left(tr \left(\Sigma^{-1} \Lambda_i \right) + \mu_i^T \Sigma^{-1} \mu_i \right) + \frac{NK}{2}
\end{aligned} \tag{4.2}
$$

Note that the multivariate longitudinal count model is embedded in the variational lower bound specified for the joint count and duration model, resulting in exactly the case discussed in Ormerod and Wand (2011).

4.3.4 Variational Approximation Estimator and Inference

Estimators for the parameters of interest can be obtained by maximizing the variational lower bound:

$$
(\hat{\beta}, \hat{\Sigma}, \hat{\gamma}, \hat{\alpha}, \hat{\mu}, \hat{\Lambda}) = \underset{(\beta, \Sigma, \gamma, \alpha, \mu, \Lambda)}{\operatorname{argmax}} \; l^*(\beta, \Sigma, \gamma, \alpha, \mu, \Lambda)
$$

This optimization now involves maximizing over the variational parameters as well as the parameters of interest. The lower bound result implies that maximizing over the variational parameters narrows the gap between the true log-likelihood, $l(\beta, \Sigma, \gamma, \alpha, \mu, \Lambda)$, and the variational log-likelihood, $l^*(\beta, \Sigma, \gamma, \alpha, \mu, \Lambda)$. Positive definiteness of Σ and Λ_i is guaranteed in the optimization algorithm by reparameterizing the variational lower bound in terms of the Cholesky decomposition of these covariance matrices with exponentiated

diagonal elements.[5]

Approximate standard errors are obtained via the Hessian matrix, H, by treating the variational lower bound as a log-likelihood. That is, approximate standard errors can be defined as the square root of the diagonal entries of the inverse observed Fisher information associated with the variational lower bound. The block diagonal structure of the variational Hessian matrix allows for efficient computation of the inverse observed Fisher information. Specifically, as described in the Appendix of Ormerod and Wand (2011), it follows from properties of matrix inversion for block diagonal matrices that the asymptotic covariance can be computed as:

$$\hat{Cov}(\hat{\beta}, \hat{\Sigma}, \hat{\gamma}, \hat{\alpha}) = -\left[H_{\theta\theta} - \sum_{i=1}^{N} H_{\theta\eta_i} H_{\eta_i\eta_i}^{-1} H_{H_{\theta\eta_i}}^{T}\right]^{-1}$$

where $\theta = [\beta, \Sigma, \gamma, \alpha]$ and $\eta_i = [\mu_i, \Lambda_i]$.

Hall et al. (2011) establish theoretical properties of the variational approximation maximum likelihood estimation for a simple Poisson mixed model. While these properties may be similarly derived for the longitudinal submodel, rigorous asymptotics for the variational approximation maximum likelihood estimation for the joint model are not presented in this work.

It is important to note the relationship between the Laplace approximation and GVA. Opper and Archambeau (2009) show the strong similarity between the two approximations. Generally, the Laplace approximation involves fitting the mean of a Gaussian density locally at the posterior maximizing point of the random effects, while GVA is a global

[5]R is used for GVA estimation and inference. Programs from Ormerod and Wand (2011) are used when applicable, specifically to estimate the measurement submodel parameters in the two-stage approach and for starting values for the joint model. The "optim" function with analytical gradient and BFGS option is used for GVA estimation of the joint model in simulation studies. Due to memory limitations, an adaptation of the Newton-Raphson algorithm described in Ormerod and Wand (2011) is used for GVA estimation of the joint model in the empirical application.

approximation where the local conditions of the Laplace approximation hold on average. Ormerod and Wand (2011) restate this relationship in the context of grouped data GLMM, where GVA can be interpreted as the Laplace approximation averaged over the assumed posterior distribution, $q(b|\mu, \Lambda)$. They show that this relationship implies that as the covariance matrix Λ goes to zero, GVA reduces to the Laplace approximation.

4.3.5 Alternative Two-Stage Estimation and Inference

We compare the GVA approach to the naive two-stage method which involves: (1) fitting the measurement submodel and estimating the unobserved heterogeneity by the empirical Bayes predictions, $E(\mathbf{b}_i|\mathbf{y}_i)$, and (2) separately fitting the duration model using the empirical Bayes estimate as a covariate. We choose the two-stage approach as a comparison because we are interested in an estimation technique that is similar in computational efficiency and ease of implementation. Two versions of the two-stage approach are used in this research: univariate and multivariate. The univariate two-stage approach ignores any and all correlation between the multiple outcomes by using empirical Bayes predictions obtained from separate univariate generalized linear mixed models as covariates in separate duration models.[6] The univariate empirical Bayes predictions, assuming Poisson and Gaussian distributions, are:

$$E(b_{ik}|y_{ik}) = \int_{b_{ik}} b_{ik} \frac{f(y_{ik}|b_{ik})f(b_{ik})}{f(y_{ik})} db_{ik} = \hat{\tau}_{ik}$$

Application of the univariate two-stage approach is simple and straightforward, but we are interested in extending the approach to multivariate longitudinal and duration

[6]The GLLAMM package in Stata (Rabe-Hesketh et al., 2005) is used for the univariate two-stage method. GLLAMM fits a Poisson random effects model using numerical integration by adaptive quadrature. Twenty integration points are used for greater accuracy.

outcomes. The multivariate two-stage approach estimates the joint multivariate longitudinal count model using GVA, and treats the approximate best predictor of the random effects, $\hat{\mu}_i$, from the measurement submodel as covariates in separate duration models. The maximizing variational parameters, $\hat{\mu}$ and $\hat{\Lambda}$ can be used as predictions for the random effects and variability associated with those predictions (Ormerod and Wand, 2011).[7] In particular, in the GVA framework:

$$E(\mathbf{b}_i|\mathbf{y}_i) = \int_{b_{iK}} ... \int_{b_{i1}} \mathbf{b}_i \phi(\mathbf{b}_i|\hat{\mu}_i, \hat{\Lambda}_i) d\mathbf{b}_i = \hat{\mu}_i$$

This best predictor is estimated based on the joint model for multivariate longitudinal count outcomes.[8]

The duration and association submodels for the two-stage approaches can be summarized as follows:

Two-Stage Duration Submodel:

$$\begin{aligned} \text{Univariate:} \quad & h(t|\mathbf{z}_{ij}, \hat{\tau}_{i1}, \ldots, \hat{\tau}_{iK}) = h_0(t) e^{\sum_{k=1}^{K} \hat{\tau}_{ik}\alpha_j} \xi_{ij} \\ \text{Multivariate:} \quad & h(t|\mathbf{z}_{ij}, \hat{\mu}_i) = h_0(t) e^{\hat{\mu}_i^T \alpha_j} \xi_{ij} \end{aligned}$$

Two-Stage Association Submodel:

$$\begin{aligned} \text{Univariate:} \quad & (b_{i1}, \ldots, b_{iK}) \sim \prod_{k=1}^{K} N(0, \sigma_k^2) \\ \text{Multivariate:} \quad & (b_{i1}, \ldots, b_{iK}) \sim MVN(0, \Sigma) \end{aligned}$$

[7] Numerical integration could also be used to obtain the empirical Bayes estimates based on the multivariate model. We find, empirically, that the multivariate variational best predictor is essentially equivalent to the multivariate empirical Bayes estimate obtained from Gauss-Hermite quadrature. In fact, in the insurance data we find that the correlation between these two estimates is about 1.0.

[8] The distinction between using best predictions of the random effects from a univariate versus a multivariate random effects count model is important and interesting in its own right. Please see Chapter 3 (Barseghyan et al., 2012) for a detailed account of the effect specific to insurance claim count data.

We may also use the methods presented in Chapter 2 and Chapter 3 to obtain the best predictor of the unobserved heterogeneity from the longitudinal submodel. In this case, a consistent estimate of Σ from the semiparametric approach for multivariate longitudinal count data is used to estimate the univariate or multivariate prediction of the unobserved heterogeneity. Just as in the two-stage approach described above, this estimate is then used as a covariate in the duration submodel. Simulation and data analysis results will be presented for this semiparametric two-stage method as well as the parametric (GLMM and GVA) two-stage method.

The two-stage approach has the advantage of simplicity and easy implementation, however it may lead to biased inference. Generally, bias results from failing to jointly incorporate the count and the duration outcomes. In particular, bias in the measurement submodel parameters is caused by ignoring the mechanism that generates the unbalanced structure of the longitudinal data thus relying on the strong assumption of missing completely at random; and bias in the duration submodel parameters results from ignoring uncertainty of estimation in the first stage. The magnitude of this bias depends on the strength of the association between the measurement and duration submodels, as well as the magnitude of the variability of the empirical Bayes estimates.

4.4 Simulation Studies

A Monte Carlo simulation study is carried out to assess feasibility and finite sample properties of the proposed estimators. We assess the bias and variability of the joint GVA approach compared to both the univariate and the multivariate two-stage approaches.

4.4.1 Simulation Design

Set the number of count outcomes $K = 3$, the number of duration outcomes $J = 2$, the number of subjects $N = 1000$, the maximum number of time periods $\max(T_i) = 9$, each $\mathbf{x}_{itk} = [1, x_{itk}]$,[9], each $\mathbf{z}_{ij} = [1]$, β_k the set consisting of an intercept parameter β_{k0} and a slope parameter β_{k1}, γ_j the set consisting of only an intercept γ_{j0}, and $\mathbf{b}_i$ the 3x1 vector of random effects distributed according to the multivariate normal with mean 0 and covariance matrix Σ. Varying levels of dependence of the K count outcomes are assumed on the J duration outcomes as measured through the parameter α. The following empirically relevant data generating process is considered:

$$y_{itk} \sim Poisson(\lambda_{itk}) \text{ with } \lambda_{itk} = e^{\beta_{k0} + x_{itk}\beta_{k1} + b_{ik}}$$

$$T^*_{ij} \sim Weibull(r_j,\ e^{b_{i1}\alpha_{j1} + b_{i2}\alpha_{j2} + b_{i3}\alpha_{j3}}\xi_{ij}) \text{ with } \xi_{ij} = e^{\gamma_{j0}}$$

$$C_i \sim Uniform(1, 9) \text{ and } T_{ij} = min(T^*_{ij}, C_i)$$

$$\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3}) \sim MVN(0, \Sigma)$$

All simulation results rely on correct distributional assumptions for the longitudinal submodel (Poisson), duration submodel (Weibull) and random effects (multivariate normal). The covariate x_{itk} is generated as draws from the empirical distribution of the insurance score variable in the insurance data sample and the simulation study parameters are similar to estimates observed in multivariate two-stage analysis of the insurance data.

[9] In general, and otherwise in the chapter, the first element of $\mathbf{x}_{itk}$ is assumed to be 1 to account for the intercept.

$$\beta_{10} = 1.24,\ \beta_{20} = .955,\ \beta_{30} = -.419$$

$$\beta_{11} = -.016,\ \beta_{21} = -.010,\ \beta_{31} = -.008$$

$$r_1 = r_2 = 2 \text{ and } \gamma_{10} = \gamma_{20} = -4.50$$

$$\alpha_1 = [0.0, .25, -.20] \text{ and } \alpha_2 = [-.20, 1.5, -.70]$$

$$\text{Scenario A: } \Sigma = \begin{bmatrix} .50 & .10 & .25 \\ .10 & .15 & .20 \\ .25 & .20 & .55 \end{bmatrix} \quad \text{Scenario B: } \Sigma = \begin{bmatrix} .50 & 0.0 & 0.0 \\ 0.0 & .15 & 0.0 \\ 0.0 & 0.0 & .55 \end{bmatrix}$$

The simulation study assumes non-random attrition, that is, dependence between the longitudinal model and the duration model. This level of dependence varies with the value of α. Levels of α are chosen to reflect what is observed in the analysis of the insurance data, with the largest positive and negative relationship in the second duration outcome for count outcomes two and three. Two simulation study scenarios are presented. Scenario A assumes a covariance structure of the random effects similar to that observed in the analysis of the insurance data. Scenario B assumes that there is no correlation between the random effects, i.e. off-diagonal elements of the random effects covariance matrix are zero.

4.4.2 Simulation Results for Parametric Estimation

The simulation studies reveal an interesting comparison of the three joint modeling methods: (1) the multivariate two-stage performs better than univariate two-stage when unobserved heterogeneity correlation is present, (2) the multivariate two-stage performs similar to univariate two-stage when unobserved heterogeneity correlation is not present,

and (3) the multivariate two-stage performs better than joint GVA in terms of bias and precision when unobserved heterogeneity correlation is present, but this advantage is lost when random effects are not correlated. Overall, we find that there is a lot to gain by jointly modeling multiple outcomes.

The properties of the association parameters, Σ and α, are the main focus of this research. Figure 4.1 presents a graphical depiction of the Monte Carlo sampling distribution of $\hat{\alpha}$, the set of parameters that measures the association between the random effects shared by the longitudinal and the duration model. Simulation scenario A, assuming positive correlation of the random effects, shows that generally the multivariate two-stage estimator has the smallest bias, with the bias of the joint GVA estimator falling in between that of the univariate and the multivariate two-stage method. The relative bias is the largest for the most extreme values of α. We find that the relative bias, $(\hat{\theta}_{MC} - \theta)/\theta$, for the case of $\alpha_{22} = 1.50$ is about $-.58$, $-.07$ and $.32$, for the univariate two-stage, multivariate two-stage and joint GVA approach, respectively. The univariate two-stage approach has the smallest Monte Carlo variance, though this precision is centered around a biased estimate. The Monte Carlo standard errors of the joint GVA estimator range from about 1.14 to 1.40 times larger than the Monte Carlo standard error of the multivariate two-stage method (see Table 4.2).

With respect to standard error estimates for $\hat{\alpha}$, we find that the univariate two-stage method results in mean estimated standard errors that are about 1.00 to 1.078 times the sampling standard deviation. The multivariate two-stage and joint GVA result in average standard error estimates that are .852 to 1.045 times and .422 to .847 times the sampling standard deviation, respectively, where the greatest underestimation is found for the largest and smallest values of α.

Figure 4.2 presents a graphical depiction of the simulation study results for estimates of Σ, the covariance matrix of the random effects. Simulation scenario A shows that generally the univariate two-stage estimator has the smallest bias, though this bias is very similar in magnitude to that of the joint GVA estimator and the multivariate two-stage method. However, the univariate two-stage approach exhibits a larger Monte Carlo standard error which results in similar levels of root mean square error for all three methods.

Simulation scenario B illustrates that in the univariate two-stage and joint GVA approaches, the bias for the largest positive and negative effects persists, albeit at a smaller magnitude, even when no correlation exists between the random effects. This suggests that the bias is likely not attributable to correlation in the posterior expectations. Scenario B also illustrates the impact of using a multivariate model in the first stage of the two-stage method. We see that when no correlation exists between the random effects, the multivariate two-step exhibits similar bias and variance as the univariate two-stage approach. In which case, the advantage of smaller bias in the multivariate two-stage approach is lost.

Table 4.2: Simulation Study Results for Parametric Estimation of Association Parameters: Σ and α

$K = 3$, $J = 2$, $N = 1000$, $\max(T_i) = 9$, 500 replications

	Scenario A						Scenario B					
Parameter	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$
Univariate GLMM Two-Stage												
σ_{11}	0.50	0.495	-0.005	0.044	0.043	0.039	0.50	0.494	-0.006	0.040	0.039	0.038
σ_{22}	0.15	0.149	-0.001	0.019	0.019	0.018	0.15	0.145	-0.005	0.019	0.019	0.019
σ_{33}	0.55	0.541	-0.009	0.062	0.061	0.060	0.55	0.523	-0.027	0.064	0.058	0.059
σ_{12}	0.10	.	.	.	.	.	0.00	.	.	.	.	.
σ_{13}	0.25	.	.	.	.	.	0.00	.	.	.	.	.
σ_{23}	0.20	.	.	.	.	.	0.00	.	.	.	.	.
α_{11}	0.00	-0.028	-0.028	0.123	0.120	0.120	0.00	-0.004	-0.004	0.111	0.111	0.114
α_{12}	0.25	0.085	-0.165	0.312	0.265	0.275	0.25	0.267	0.017	0.249	0.248	0.262
α_{13}	-0.20	-0.103	0.097	0.169	0.138	0.143	-0.20	-0.201	-0.001	0.129	0.129	0.136
α_{21}	-0.20	-0.216	-0.016	0.120	0.118	0.120	-0.20	-0.163	0.037	0.108	0.102	0.111
α_{22}	1.50	0.636	-0.864	0.899	0.248	0.267	1.50	1.176	-0.324	0.399	0.233	0.244
α_{23}	-0.70	-0.256	0.444	0.466	0.140	0.143	-0.70	-0.566	0.134	0.185	0.129	0.140
Multivariate GVA Two-Stage												
σ_{11}	0.50	0.480	-0.020	0.043	0.038	.	0.50	0.480	-0.020	0.043	0.038	.
σ_{22}	0.15	0.146	-0.004	0.019	0.018	.	0.15	0.142	-0.008	0.020	0.018	.
σ_{33}	0.55	0.519	-0.031	0.063	0.055	.	0.55	0.491	-0.059	0.079	0.052	.
σ_{12}	0.10	0.099	-0.001	0.019	0.019	.	0.00	0.001	0.001	0.019	0.019	.
σ_{13}	0.25	0.244	-0.006	0.034	0.033	.	0.00	-0.004	-0.004	0.033	0.033	.
σ_{23}	0.20	0.197	-0.003	0.024	0.024	.	0.00	0.010	0.010	0.026	0.024	.
α_{11}	0.00	0.002	0.002	0.149	0.149	0.156	0.00	-0.004	-0.004	0.114	0.114	0.116
α_{12}	0.25	0.305	0.055	0.575	0.573	0.575	0.25	0.282	0.032	0.258	0.256	0.270
α_{13}	-0.20	-0.218	-0.018	0.316	0.316	0.324	-0.20	-0.211	-0.011	0.135	0.134	0.143
α_{21}	-0.20	-0.171	0.029	0.163	0.160	0.157	-0.20	-0.169	0.031	0.115	0.111	0.114
α_{22}	1.50	1.391	-0.109	0.672	0.663	0.565	1.50	1.222	-0.278	0.374	0.251	0.255
α_{23}	-0.70	-0.670	0.030	0.373	0.372	0.325	-0.70	-0.599	0.101	0.173	0.140	0.148
Joint GVA												
σ_{11}	0.50	0.484	-0.016	0.041	0.038	.	0.50	0.483	-0.017	0.042	0.038	.
σ_{22}	0.15	0.146	-0.004	0.019	0.018	.	0.15	0.145	-0.005	0.019	0.018	.
σ_{33}	0.55	0.526	-0.024	0.061	0.056	.	0.55	0.517	-0.033	0.064	0.055	.
σ_{12}	0.10	0.099	-0.001	0.019	0.019	.	0.00	-0.001	-0.001	0.019	0.019	.
σ_{13}	0.25	0.248	-0.002	0.034	0.034	.	0.00	-0.002	-0.002	0.034	0.034	.

Continued on next page...

	Scenario A						Scenario B					
Parameter	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\overline{\hat{se}(\hat{\theta})}$
σ_{23}	0.20	0.197	-0.003	0.025	0.024	.	0.00	0.003	0.003	0.024	0.024	.
α_{11}	0.00	0.017	0.017	0.184	0.183	0.155	0.00	0.002	0.002	0.132	0.132	0.118
α_{12}	0.25	0.345	0.095	0.691	0.684	0.425	0.25	0.255	0.005	0.299	0.299	0.265
α_{13}	-0.20	-0.259	-0.059	0.389	0.385	0.253	-0.20	-0.222	-0.022	0.161	0.159	0.140
α_{21}	-0.20	-0.189	0.011	0.219	0.218	0.134	-0.20	-0.219	-0.019	0.147	0.146	0.134
α_{22}	1.50	1.983	0.483	0.983	0.856	0.361	1.50	1.612	0.112	0.348	0.330	0.525
α_{23}	-0.70	-0.966	-0.266	0.559	0.492	0.210	-0.70	-0.847	-0.147	0.241	0.191	0.247

Note: $\hat{\theta}_{MC} = .002 \sum_{r=1}^{500} \hat{\theta}^{(r)}$ is the Monte Carlo estimate of the parameter, bias is the difference between the true value θ and $\hat{\theta}_{MC}$, $RMSE = \sqrt{.002 \sum_{r=1}^{500} \left(\hat{\theta}^{(r)} - \theta\right)^2}$ is the root mean squared error, $se(\hat{\theta})$ is the Monte Carlo standard deviation of $\hat{\theta}$, $\overline{\hat{se}(\hat{\theta})}$ is the mean of the estimated standard error.

4.4.3 Simulation Results for Semiparametric Estimation

We are also interested in the performance of the univariate and multivariate two-stage approach in which the first stage is estimated semiparametrically. Table 4.3 and Figure 4.3 display the simulation study results for the association parameters, α, using the parametric and semiparametric two-stage methods and data as simulated in Scenario A. Generally, we find that the empirical MSE associated with the estimation of the association parameters using the semiparametric two-stage method is about 1.2 and 2.4 larger than the empirical MSE associated with parametric estimation. Note that the simulated data is correctly specified in the parametric approach, i.e. the random effects are normally distributed. In this case, we expect the semparametric approach to be less precise; however, in misspecified cases we expect the semiparametric approach to perform better as in Chapter 2.

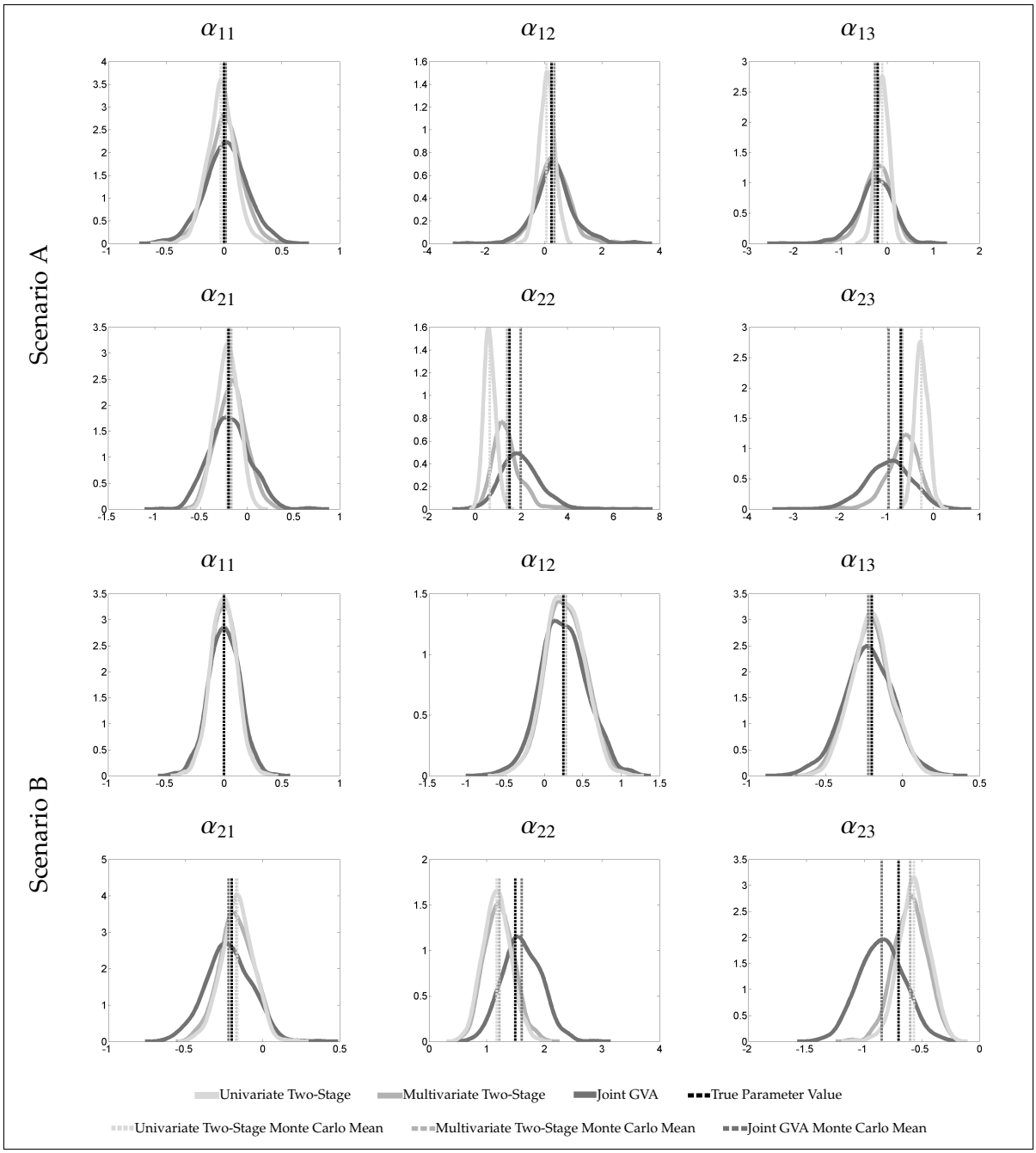

Figure 4.1: Kernel Density Plots of Random Effects Coefficient Parameter Estimates from Simulation Study: Parametric

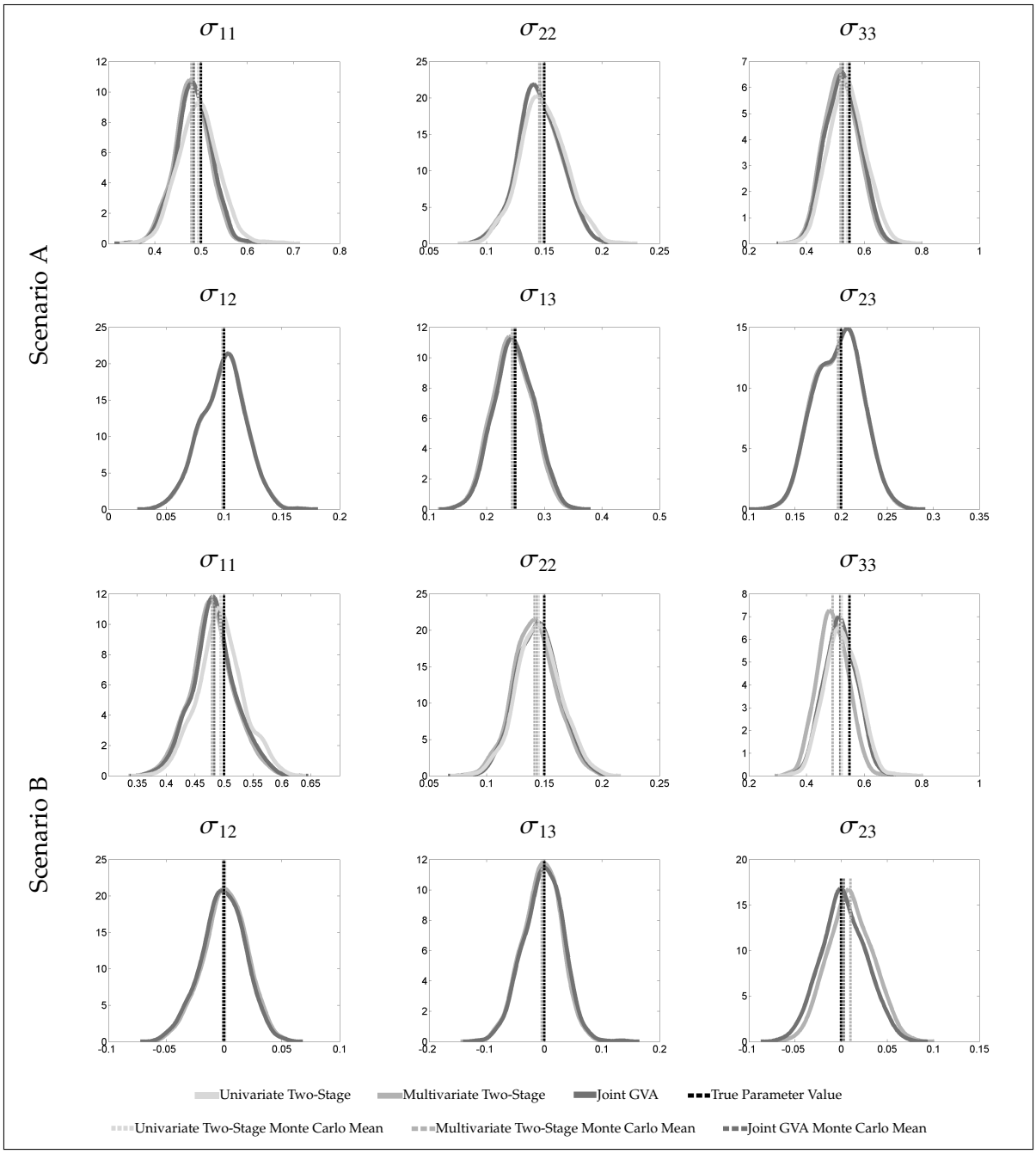

Figure 4.2: Kernel Density Plots of Variance/Covariance Parameter Estimates from Simulation Study: Parametric

Table 4.3: Simulation Study Results for Semiparametric and Parametric Two-Stage Estimation of Random Effects Coefficient, α

$K = 3$, $J = 2$, $N = 1000$, $\max(T_i) = 9$, 500 replications, Scenario A

		GLMM/GVA				Semiparametric			
Parameter	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$
Univariate Two-Stage									
α_{11}	0.00	-0.028	-0.028	0.123	0.120	0.053	0.053	0.166	0.157
α_{12}	0.25	0.085	-0.165	0.312	0.265	0.144	-0.106	0.482	0.470
α_{13}	-0.20	-0.103	0.097	0.169	0.138	-0.071	0.129	0.244	0.207
α_{21}	-0.20	-0.216	-0.016	0.120	0.118	-0.187	0.013	0.158	0.158
α_{22}	1.50	0.636	-0.864	0.899	0.248	1.095	-0.405	0.620	0.469
α_{23}	-0.70	-0.256	0.444	0.466	0.140	-0.299	0.401	0.458	0.221
Multivariate Two-Stage									
α_{11}	0.00	0.002	0.002	0.149	0.149	0.058	0.058	0.191	0.182
α_{12}	0.25	0.305	0.055	0.575	0.573	0.184	-0.066	0.730	0.727
α_{13}	-0.20	-0.218	-0.018	0.316	0.316	-0.109	0.091	0.352	0.340
α_{21}	-0.20	-0.171	0.029	0.163	0.160	-0.200	0.000	0.207	0.207
α_{22}	1.50	1.391	-0.109	0.672	0.663	1.561	0.061	1.002	1.000
α_{23}	-0.70	-0.670	0.030	0.373	0.372	-0.565	0.135	0.486	0.467

Note: $\hat{\theta}_{MC} = .002 \sum_{r=1}^{500} \hat{\theta}^{(r)}$ is the Monte Carlo estimate of the parameter, bias is the difference between the true value θ and $\hat{\theta}_{MC}$, $RMSE = \sqrt{.002 \sum_{r=1}^{500} \left(\hat{\theta}^{(r)} - \theta\right)^2}$ is the root mean squared error, $se(\hat{\theta})$ is the Monte Carlo standard deviation of $\hat{\theta}$.

4.5 An Empirical Application: Insurance Data

4.5.1 Empirical Results

All two-stage methods and the joint model GVA approach are used to analyze the insurance data described in Section 4.2. We find that the unobserved risk component has statistically significant power in explaining the duration of auto policies in our sample,

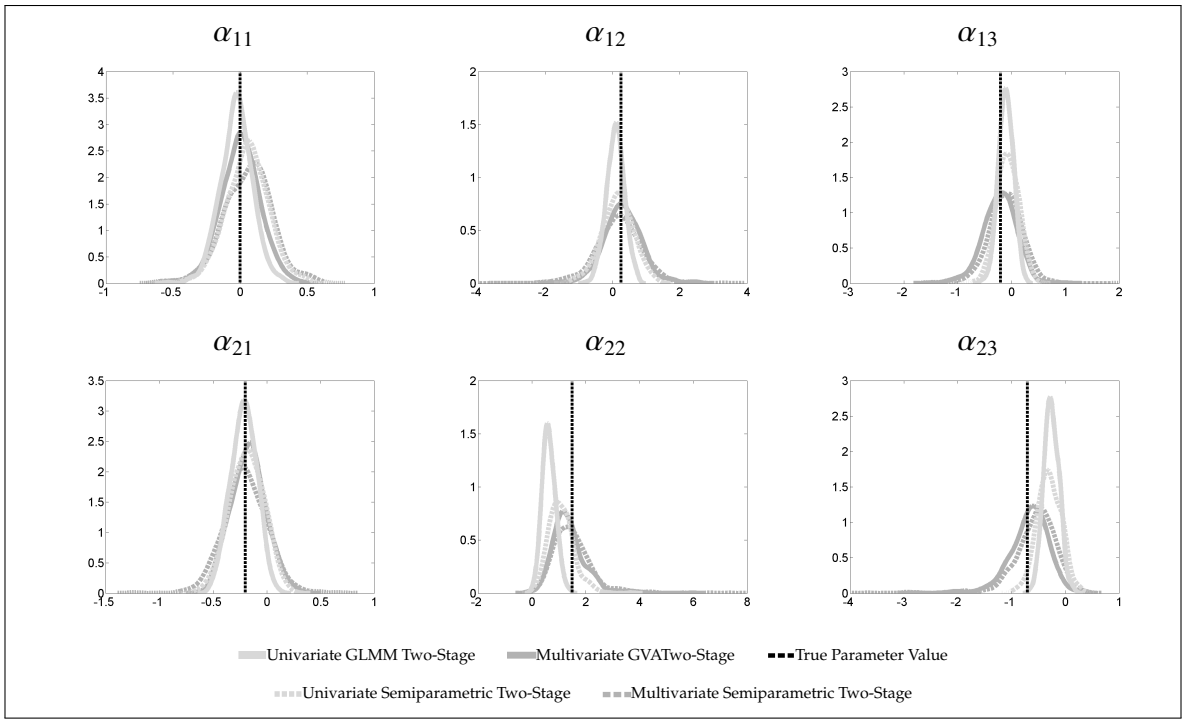

Figure 4.3: Kernel Density Plots of Random Effects Coefficient Parameter Estimates from Simulation Study: Semiparametric and Parametric Two-Stage

suggesting that economic forces are at play in shaping the overall relation between the company and its clients.

Dependence between the unobserved "riskiness" of a policyholder and propensity to maintain policies in force can be assessed through the parameter α. In this analysis, we find that the joint GVA and multivariate two-stage approach give roughly similar estimates of this relationship (see Table 4.4), particularly in the analysis of home duration. In the auto duration model, these two methods estimate the same direction of the effect and capture the large magnitude of the effect of unobserved heterogeneity measured through collision claims on the auto policy hazard. The GVA multivariate two-stage and joint GVA method find that the hazard is roughly 10 times larger for a unit increase in unob-

served "riskiness" as measured through collision claims. This effect is estimated to be slightly smaller using the multivariate semiparametric two-stage approach. The univariate two-stage approach estimates a positive relation between collision claim unobserved heterogeneity and auto duration hazard, though the effect is about half the size. The semiparametric and parametric two-stage approaches find roughly similar effects in the univariate and the multivariate case. However, the semiparametric two-stage method estimates larger standard errors resulting in fewer statistically significant results.

Table 4.4: Association Parameter Results for Analysis of Insurance Claim Data

	Univariate Two-Stage				Multivariate Two-Stage					
Parameter	GLMM		Semipar.		GVA		Semipar.		Joint GVA	
Home: e^{α_H}	1.10∗	(.045)	1.01	(.097)	1.00	(.072)	1.19	(.473)	.980	(.121)
Home: e^{α_L}	1.20	(.133)	1.13	(.160)	1.33	(.387)	1.32	(.473)	1.27	(.864)
Home: e^{α_P}	.951	(.053)	.984	(.097)	.815	(.133)	.790	(.347)	.816	(.329)
Auto: e^{α_H}	.929	(.038)	.858	(.080)	.822∗	(.057)	.918	(.348)	.714∗	(.086)
Auto: e^{α_L}	4.82∗	(.468)	4.67∗	(.598)	10.93∗	(2.90)	7.98∗	(2.64)	9.95∗	(2.94)
Auto: e^{α_P}	1.14∗	(.059)	1.18	(.110)	.502∗	(.076)	.595	(.247)	.630∗	(.128)

Notes: Includes 27,051 Policies, 98,804 Observations. Standard errors are in parentheses. * indicates statistical significance at the 5% level.

It is important to note the slight differences in the estimates of the elements of the covariance matrix Σ from the multivariate two-stage and the joint GVA methods (see below). This implies that the duration outcomes are making a contribution to the estimation of the association between the unobserved heterogeneity for the three types of coverage.

$$\text{Univ. GLMM Two-Stage: } \hat{\Sigma} = \begin{bmatrix} .558 & . & . \\ . & .165 & . \\ . & . & .637 \end{bmatrix} \quad \text{Joint, Home: } \hat{\Sigma} = \begin{bmatrix} .511 & .089 & .259 \\ .089 & .161 & .204 \\ .259 & .204 & .564 \end{bmatrix}$$

$$\text{Multiv. GVA Two-Stage: } \hat{\Sigma} = \begin{bmatrix} .498 & .093 & .255 \\ .093 & .141 & .193 \\ .255 & .193 & .549 \end{bmatrix} \quad \text{Joint, Auto: } \hat{\Sigma} = \begin{bmatrix} .513 & .096 & .265 \\ .096 & .148 & .202 \\ .265 & .202 & .568 \end{bmatrix}$$

In this data, there exists a strong positive correlation between the two duration outcomes since policy cancellation decisions for auto and home coverages are often made jointly. Unfortunately, the fully joint model that estimates the three claim count outcomes and two duration outcomes together, did not attain reasonable convergence. Thus results presented here do not jointly incorporate the two duration outcomes, rather two separate models are estimated: (1) trivariate count outcomes with home duration and (2) trivariate count outcomes with auto duration. The strong correlation may imply the need for a model that allows the multiple durations to be related in a more general way, specifically one that captures the relationship between durations through additional random effects. Such alternative models are discussed in Section 6.1. Investigation of the effect of strongly correlated duration outcomes is on-going.

Each univariate generalized linear mixed model (GLMM) takes about 8 hours to estimate for a total of about 24 hours to obtain the best predictors via numerical integration in the univariate two-stage approach. The second stage fitting of the Weibull proportional hazards model has trivial computing time. For the multivariate two-stage approach, the GVA method used to obtain best predictors for the random effects takes about 9 hours. Estimates from the semiparametric two-stage approaches are obtained in less than 2 hours.

The fitting of the multivariate longitudinal count model is computationally prohibitive with standard numerical integration (Morris, 2011). These results illustrate the computational improvement obtained by using GVA or the semiparametric approach from Chapter 2 for estimating the first stage of the two-stage approach rather than numerical integration techniques. The joint GVA takes about 24 hours to attain convergence.

4.5.2 Posterior Expectation of Unobserved Heterogeneity

The posterior expectation of the random effects are important quantities of interest. In the case of two-stage estimation, these best predictors are the quantities that are plugged into the second stage duration model. Figure 4.4 presents kernel density plots of the posterior expectations estimated from the univariate GLMM two-stage, multivariate GVA two-stage and joint GVA methods. We find that the posterior means estimated from the multivariate GVA two-stage and joint GVA approach display a smoother distribution than those estimated from the univariate GLMM two-stage approach. In the case of comprehensive claims, where we observe the smallest claim rate and thus the least variation in count outcome, we find a very pronounced difference in the posterior mean distribution for the univariate GLMM two-stage approach versus the multivariate approaches. This suggests an advantage for using multivariate information in estimating unobserved heterogeneity. Please see Chapter 3 (Barseghyan et al., 2012) for a thorough study and in depth discussion of the benefit of using multivariate modeling with respect to posterior expectations. While generally the posterior means estimated from the multivariate GVA two-stage and the joint GVA methods are similar, we find that the posterior expectation associated with collision claims behaves slightly differently. For this case, the distribution of posterior expectation is substantially smoother for the joint GVA model of the trivari-

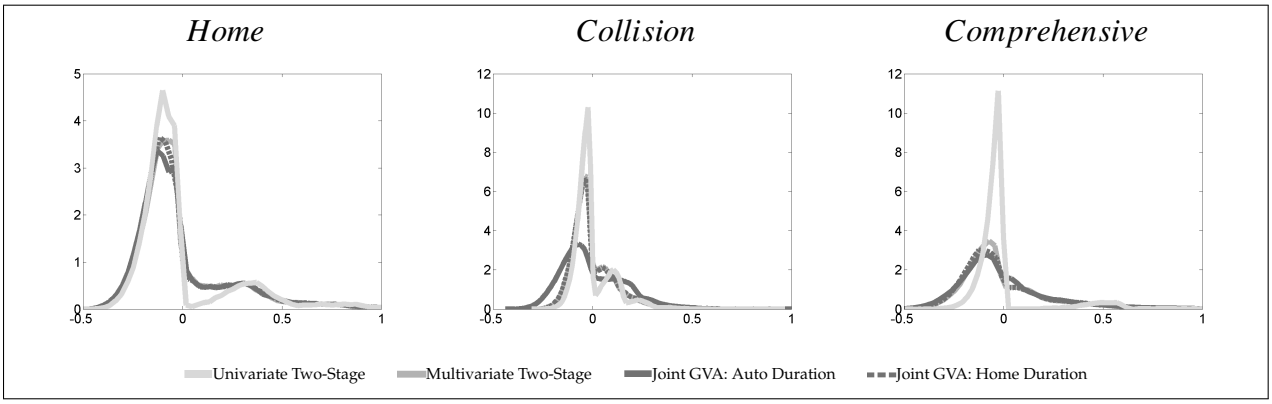

Figure 4.4: Kernel Density Plots of Estimated Posterior Means from Insurance Data

ate count outcomes and auto duration, suggesting that observed auto duration strongly influences the posterior mean. This result is consistent with the statistical and economic significance of the collision unobserved heterogeneity coefficient in the joint model.

4.6 Discussion

4.6.1 Alternate Joint Random Effects Models

The model described in this paper incorporates unobserved heterogeneity in the duration model and association between duration and measurement models through a specific structure determined by the parameter α. Many alternate random effect models can be specified to account for the relationship between the measurement and duration models through other dependencies. For example, the econometrics literature focuses on "Multivariate Mixed Proportional Hazard " (MMPH) models for multiple duration outcomes (Heckman and Leemer, 2001). In the MMPH model specification, each marginal duration

distribution follows a univariate proportional hazards random effects model and the dependence between multiple durations is imposed through correlation in the unobserved heterogeneity. The MMPH models can be extended to multivariate joint models by allowing the duration unobserved heterogeneity to also be correlated with the longitudinal unobserved heterogeneity:

Alternate Measurement Submodel:

$$y_{itk}|\mathbf{x}_{itk}, b_{ik} = e^{b_{ik}}\lambda_{ik}$$

Alternate Duration Submodel:

$$h(T^*_{ij}|\mathbf{z}_{ij}, \mathbf{b}_i) = h_0(T^*_{ij})e^{\nu_{ij}}\xi_{ij}$$

Alternate Random Effects Specification:

$$f_i([\mathbf{b}_i, \nu_i]|\Sigma) \sim G_{K+J}(0, \Sigma)$$

where G is a $(K + J)$-dimensional multivariate distribution. Note that this model replaces the $\mathbf{b}_i^T \alpha_j$ term in the duration submodel for the joint model proposed in Section 3.1 with ν_i, which implies a more general association between the measurement and duration submodels. Such models are likely good candidates for Gaussian variational approximation. Future research will investigate the most appropriate model for our economic question of interest.

4.6.2 Variational Parameters: μ_i and Λ_i

GVA estimation of the joint longitudinal and duration model may be prone to convergence issues concerning the identifiability of the variational parameters. As discussed in

Section 4.3.5, the model is specified such that μ_i and Λ_i act as the mean and covariance matrix of the assumed approximate posterior distribution of unobserved heterogeneity. That is, the appropriate approximation to the best predictor integral is one that replaces the true posterior distribution with the assumed posterior distribution evaluated at the maximizing variational parameters $\hat{\mu}_i$ and $\hat{\Lambda}_i$ (Ormerod and Wand, 2011). To investigate potential identification issues, we carry out a simulation study using the same data generating process as in Section 4.4, but only include the subject's first count observation in the estimation of the count model. With less information per subject, i.e. only one time period of count information, we find that convergence is not attained in about 17% of the simulation repetitions. Such is the case for both the multivariate GVA two-stage and joint GVA methods. In the application of these methods to the insurance data, the $4,748$ policies (which represents about 17.5% of the policies) observed for only one time period may be a contributing factor to the convergence issues.

In the simulation study with one time period of count information, when convergence is achieved[10] we find that the bias for the random effects coefficient parameters is much larger than the corresponding all time period simulation study for both the multivariate GVA two-stage and joint GVA approaches. The empirical bias for the multivariate GVA two-stage and joint GVA approach is as much as about 22 and 4 times larger, respectively, in the one time period simulation study compared to the all time periods simulation study. Furthermore, the empirical RMSE from the one time period simulation study is about 8 to 11.5 times greater for the multivariate GVA two-stage and about 3.5 to 6.2 times greater for the joint GVA approach than that found in the all time periods simulation study. The univariate GLMM two-stage approach is robust to the bias, though it

[10]Convergence in the GVA method is attained when the following quantities are close to zero: (1) percent change in model parameters $\hat{\theta}$, and (2) average square of gradient vector with respect to model parameters θ.

exhibits a larger Monte Carlo variance in the one time period simulation study, making the estimator about 1.1 to 3.1 times more efficient in the all time period case. The simulation studies indicate that joint GVA is more robust to limited count information than the multivariate GVA two-stage approach. The empirical MSE of the multivariate GVA two-stage method is smaller than the empirical MSE of the joint GVA method in the all time periods simulation study, but the opposite is true in the one time period simulation study.

Table 4.5: Simulation Study Results for One Time Period Parametric Estimation of Random Effects Coefficient, α

$K = 3$, $J = 2$, $N = 1000$, $\max(T_i) = 9$ or 1, 248 replications, Scenario A

		One Time Period				All Time Periods			
Parameter	Truth	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$	$\hat{\theta}_{MC}$	Bias	RMSE	$se(\hat{\theta})$
Univariate GLMM Two-Stage									
α_{11}	0.00	-0.028	-0.028	0.179	0.177	-0.029	-0.029	0.127	0.123
α_{12}	0.25	0.026	-0.224	0.523	0.472	0.096	-0.154	0.298	0.255
α_{13}	-0.20	-0.102	0.098	0.255	0.236	-0.101	0.099	0.165	0.132
α_{21}	-0.20	-0.238	-0.038	0.181	0.177	-0.213	-0.013	0.121	0.120
α_{22}	1.50	0.577	-0.923	1.014	0.420	0.649	-0.851	0.884	0.242
α_{23}	-0.70	-0.259	0.441	0.499	0.233	-0.254	0.446	0.468	0.143
Multivariate GVA Two-Stage									
α_{11}	0.00	0.090	0.090	1.216	1.213	0.001	0.001	0.151	0.151
α_{12}	0.25	0.680	0.430	4.212	4.190	0.310	0.060	0.533	0.530
α_{13}	-0.20	-0.509	-0.309	2.787	2.769	-0.214	-0.014	0.300	0.300
α_{21}	-0.20	0.027	0.227	1.734	1.719	-0.165	0.035	0.172	0.168
α_{22}	1.50	3.104	1.604	5.757	5.529	1.374	-0.126	0.591	0.577
α_{23}	-0.70	-1.709	-1.009	3.949	3.818	-0.655	0.045	0.343	0.340
Joint GVA									
α_{11}	0.00	0.066	0.066	0.671	0.668	0.016	0.016	0.183	0.182
α_{12}	0.25	0.694	0.444	2.741	2.705	0.371	0.121	0.667	0.655
α_{13}	-0.20	-0.432	-0.232	1.498	1.480	-0.264	-0.064	0.386	0.381
α_{21}	-0.20	-0.112	0.088	1.417	1.414	-0.180	0.020	0.228	0.227
α_{22}	1.50	2.323	0.823	3.170	3.062	1.971	0.471	0.900	0.767
α_{23}	-0.70	-1.123	-0.423	1.839	1.790	-0.955	-0.255	0.527	0.461

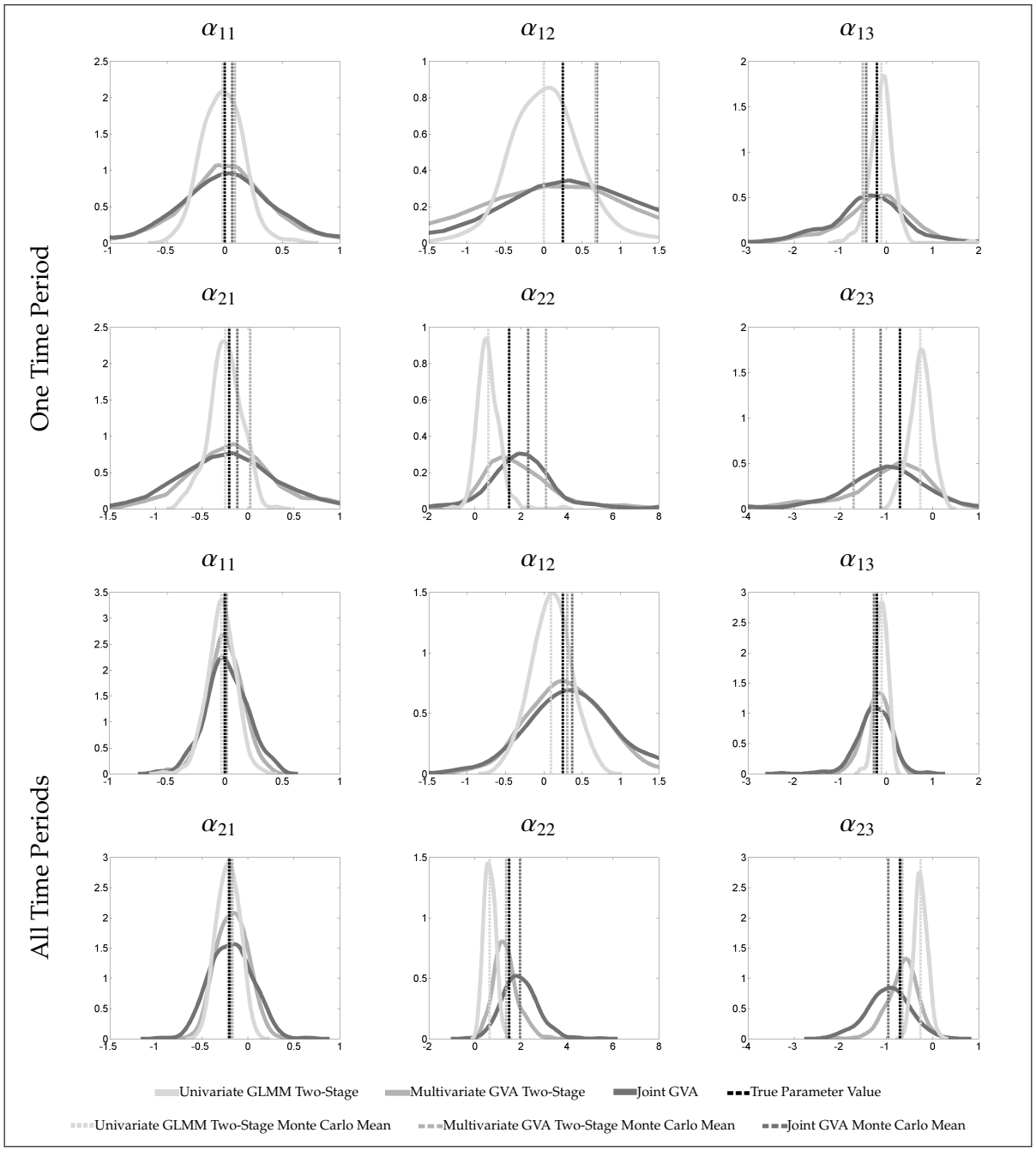

Figure 4.5: Kernel Density Plots of Random Effects Coefficient Parameter Estimates from One Time Period Simulation Study: Parametric

Note: 300 replications were run, which resulted in 248 attaining convergence. Quantities are defined as in Table 4.3.

4.6.3 Computational Advantages

The methods presented in this paper have the potential to exhibit significant computational advantage over maximum likelihood estimation of the joint longitudinal and duration model. The two-stage approaches decompose the marginal likelihood into two separate models that are straightforward to estimate, and the GVA approach circumvents a possibly high-dimensional integral by introducing variational parameters. However, it is important to point out that as the number of subjects increases, the number of variational parameters in the GVA method increases by a factor depending on the number of longitudinal outcomes, K. For example, the trivariate count outcome in the insurance data implies three posterior mean parameters and six posterior variance-covariance parameters for each subject. Thus there is a trade-off between computational complexity due to a high-dimensional integral and optimization over many parameters, particularly as the number of subjects increases.

It is computationally advantageous to perform model selection on the measurement and duration submodels separately, and then specify the link between the two models. When model selection is of interest, the two-stage approaches offer a computational advantage. Best predictors from the first-stage fitting of the multivariate longitudinal model can be easily stored and used in various specifications of the second-stage duration model, thus avoiding repeated fitting of the first-stage measurement model.

4.7 Conclusion

This chapter presents a flexible model for joint modeling of longitudinal and duration outcomes based on correlated random effects. A joint model provides a way to characterize the relationship between count and duration outcomes and to account for complications from dropout associated with the longitudinal model. Maximum likelihood estimation of such models present computational complexities, particularly when extending to non-normal longitudinal measurements and/or multivariate data. Gaussian variational approximation is a technique proposed by Ormerod and Wand (2011) as a fast, deterministic alternative to MCMC for intractable calculus problems. This paper proposes GVA as an alternative for dealing with intractable multivariate integrals in the joint multivariate longitudinal and multivariate duration model. For such a model, GVA results in a computationally tractable optimization problem, though it relies on additional assumptions regarding the subject-specific random effects.

Our simulation studies provide evidence of the importance of using multivariate information when association between multivariate longitudinal counts and multivariate duration exist. Generally, we find that multivariate two-stage estimation of the joint model is computationally very simple and performs relatively well in finite samples, while a univariate two-stage approach lacks desirable properties. The GVA approach exhibits slightly more desirable properties than the multivariate two-stage approach when no association in unobserved heterogeneity is present.

Applying the multivariate two-stage and joint GVA approach to the insurance data, we find statistically significant effects of intrinsic policyholder "riskiness" on the length of time an auto policy is maintained in force. The association between auto duration and unobserved heterogeneity, as measured through collision claims, is the strongest. The

joint GVA approach accounts for this dependence in the estimation of the association parameters of the multivariate longitudinal count model. In the same vein, we find that the joint GVA posterior expectations of unobserved heterogeneity, as measured through collision claims, exhibit a smoother empirical distribution than estimates obtained from the two-stage method. Since the posterior expectations from the joint GVA depend on observed claim counts as well as observed policy duration, this suggests that incorporating duration information that is significantly associated with unobserved heterogeneity has significant impact on estimates of unobserved heterogeneity. Incorporating duration information in posterior expectations would be computationally prohibitive with numerical techniques and inexecutable with two-stage methods by design.

APPENDIX A

CHAPTER 2 APPENDIX

Following the notation in Wooldridge (2001), $U(\beta)$ and $U(\Sigma^*)$ can be written in the framework of a general two-step M-estimator. Proofs of standard consistency and asymptotic normality results for two-step M-estimators are not presented here; rather a summary of the results and regularity conditions applicable to the semiparametric methodology for multivariate longitudinal count data is outlined. The two-step M-estimation problem solves

$$min_{\theta \in \Theta} \, N^{-1} \sum_{i=1}^{N} q(y_i, \theta; \hat{\gamma}) \tag{A.1}$$

Here $\hat{\gamma}$ is a consistent estimator for some parameter $\widetilde{\gamma}$ where $\hat{\gamma}$ may not converge to γ_0, a parameter indexing some interesting feature of the distribution. In the estimation of the regression parameters $\theta = \beta$ and $\hat{\gamma} = \hat{\mathbf{\Sigma}}^*$. In the estimation of the association parameters $\theta = \mathbf{\Sigma}^*$ and $\hat{\gamma} = \hat{\beta}$. Letting Θ be a subset of $\Re^p$ and $q : \mathcal{Y} \text{ x } \Theta \rightarrow \Re$ be a real-valued function, the following are the regularity conditions that imply the uniform weak law of large numbers.

Condition A.0.1 (Regularity Conditions for Consistency)

(*i*) Θ *is compact*

(*ii*) *for each* $\theta \in \Theta$, $q(., \theta; \hat{\gamma})$ *is Borel measurable on* $\mathcal{Y}$

(*iii*) *for each* $y \in \mathcal{Y}$, $q(y, .; \hat{\gamma})$ *is continuous on* Θ

(*iv*) $|q(y, \theta; \hat{\gamma})| \leq b(y)$ *for all* $\theta \in \Theta$ *where b is a non-negative function of* $\mathcal{Y}$ *subject to* $E(b(y)) < \infty$

Conditions A.0.1*(i)* and *(ii)* are assumed. The objective function is well-behaved in the sense that it is a product of well-behaved continuous functions on Θ. The absolute value

of the objective function is bounded by a product of functions of y that have finite expectation by the assumed moment assumptions of the model. Since $q(y_i, \theta; \hat{\gamma})$ satisfies these regularity conditions, the objective function in equation A.1 converges to $E(q(y_i, \theta; \widetilde{\gamma}))$ uniformly over Θ, i.e. the uniform weak law of large numbers holds. Consistency results also require identification conditions.

Condition A.0.2 (Identification Condition for Consistency)

$$E\left[q(y, \theta_0; \widetilde{\gamma})\right] < E\left[q(y, \theta; \widetilde{\gamma})\right] \text{ , all } \theta \in \Theta, \theta \neq \theta_0$$

where θ_0 is the true value. In the case of $U(\beta)$, this condition holds when there is no perfect collinearity in the covariates. In this case, the estimator is a special case of nonlinear least squares with an exponential mean function. In the case of $U(\mathbf{\Sigma}^*)$, this condition is satisfied by construction of the second set of estimating equations, i.e. the design matrix for subject i is the vector of unique elements of $\lambda_i \lambda_i^T$.

Consistency results rely on the assumptions of the moments presented in Result 2.3.2*(i)* and *(ii)*. In the case of $U(\beta)$, if Result 2.3.2*(ii)* is correctly specified then an estimator such that $\mathbf{\Sigma}_0^* = plim\ \hat{\mathbf{\Sigma}}^*$ can be chosen where $\mathbf{\Sigma}_0^*$ indexes the marginal variance. The consistency of the estimator obtained from solving $U(\beta) = 0$ is robust to misspecification of Result 2.3.2*(ii)* since $\hat{\mathbf{\Sigma}}^*$ is also well-defined in the case that $\widetilde{\mathbf{\Sigma}}^* = plim\ \hat{\mathbf{\Sigma}}^*$, where $\widetilde{\mathbf{\Sigma}}^*$ does not necessarily characterize some interesting feature of the distribution. Note that this is a special case of weighted nonlinear least squares. On the other hand, consistency of the estimator obtained from solving $U(\mathbf{\Sigma}^*) = 0$ holds only for $\widetilde{\beta} = \beta_0$ where β_0 indexes the marginal mean, i.e. Result 2.3.2*(i)* must hold in addition to Result 2.3.2*(ii)*.

In addition to Condition A.0.1, the following regularity conditions must be satisfied for asymptotic normality.

Condition A.0.3 (Regularity Conditions for Asymptotic Normality)

(*i*) θ_0 *is in the interior of* Θ

(*ii*) $s(y, .; \hat{\gamma})$ *is continuously differentiable on the interior of* Θ *for all* $y \in \mathcal{Y}$

(*iii*) *Each element of* $H(y, \theta; \hat{\gamma})$ *is bounded in absolute value by the function* $b(y)$ *where* $E(b(y)) < \infty$

(*iv*) $E(s(y, \theta_0; \hat{\gamma})) = 0$

(*v*) $E(H(y, \theta_0; \hat{\gamma}))$ *is positive definite*

(*vi*) *Each element of* $s(y, \theta_0; \hat{\gamma})$ *has finite second moment*

where $s(y, \theta; \hat{\gamma})$ is the derivative of the objective function and $H(y, \theta; \hat{\gamma})$ is the Hessian of the objective function. Condition A.0.3*(i)* is assumed. The score function is the product of well-defined, continuously differentiable functions over Θ. The Hessian is comprised of the matrices A, B and C defined in Result 2.3.4 that have nice properties by the assumptions of the moment conditions. These mild regularity conditions are satisfied by the assumed moment conditions and the exponential and linear form of $\lambda_i(\beta)$ and $\mathbf{V}_i^*(\hat{\beta}, \mathbf{\Sigma}^*)$, respectively.

APPENDIX B

CHAPTER 3 APPENDIX

Table B.1: Distribution of Claim Counts

Tricoverage Sample (294,917 household-years)

	Collision		Comprehensive		Home	
Count	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	265,692	90.09	285,923	96.95	273,984	92.90
1	27,186	9.22	8,495	2.88	18,886	6.40
2	1,890	0.64	467	0.16	1,872	0.63
3	140	0.05	30	0.01	159	0.05
4	6	-	0	-	12	-
5	3	-	2	-	2	-
6					2	-

Note: Dash indicates less then 0.01 percent.

Table B.2: Distribution of Deductibles

Tricoverage Sample (294,917 household-years)

	Collision		Comprehensive		Home	
Deductible	Frequency	Percent	Frequency	Percent	Frequency	Percent
50			34,007	11.53		
100	7,846	2.66	18,502	6.27	11,577	3.93
200	65,672	22.27	128,599	43.61		
250	51,644	17.51	31,556	10.70	197,100	66.83
500	159,702	54.15	78,098	26.48	70,567	23.93
1,000	10,053	3.41	4,155	1.41	14,537	4.93
2,500					1,044	0.35
5,000					92	0.03

Note: Amounts in dollars.

Table B.3: Descriptive Statistics of Premiums

Tricoverage Sample (294,917 household-years)

	Collision	Comprehensive	Home
Mean	200	127	548
Standard deviation	103	70	309
Minimum	20	6	50
1st percentile	60	34	204
5th percentile	82	48	265
10th percentile	97	58	296
25th percentile	129	81	359
Median	178	113	466
75th percentile	243	157	638
90th percentile	327	210	891
95th percentile	393	250	1,110
99th percentile	560	358	1,683
Maximum	2,520	2,524	10,224

Note: Amounts in dollars.

Table B.4: Descriptive Statistics of Covariates

Tricoverage Sample (294,917 household-years)

	Mean	Std. Dev.	Minimum	Maximum
Auto:				
Driver 1 age (years)	56.10	14.70	19	99
Driver 1 female	0.33	0.47	0	1
Driver 1 single	0.22	0.41	0	1
Driver 1 married	0.63	0.48	0	1
Driver 1 insurance score	789.51	106.50	297	996
Driver 2	0.48	0.50	0	1
Driver 2 age (years)	50.28	12.93	16	94
Driver 2 female	0.91	0.28	0	1
Driver 3+	0.04	0.21	0	1
Young driver	0.01	0.10	0	1
Vehicle 1 age (years)	4.43	3.59	-1	46
Vehicle 1 personal use	0.47	0.50	0	1
Vehicle 1 passive restraint	0.99	0.10	0	1
Vehicle 1 anti-theft	0.57	0.49	0	1
Vehicle 1 anti-lock brakes	0.79	0.41	0	1
Vehicle 2	0.53	0.50	0	1
Vehicle 2 age (years)	5.94	5.53	-1	83
Vehicle 2 personal use	0.55	0.50	0	1
Vehicle 2 passive restraint	0.94	0.24	0	1
Vehicle 2 anti-theft	0.46	0.50	0	1
Vehicle 2 anti-lock brakes	0.70	0.46	0	1
Vehicle 3+	0.05	0.22	0	1
Home:				
Home age (years)	45.05	27.20	0	206
Insured value (thousands of dollars)	153.31	75.63	1	3250
Farm or business	0.02	0.15	0	1
Primary residence	1.00	0.04	0	1
Owner occupied	0.98	0.14	0	1
Number of families	1.16	1.89	1	99
Masonry construction	0.07	0.25	0	1
Distance to fire hydrant (feet)	401.83	514.82	0	30,000
Alarm or other protection	0.95	0.22	0	1

Note: Insurance score is based on information contained in credit reports.

Table B.5: Regression Parameter Estimates - Auto

Tricoverage Sample (294,917 household-years)

	Collision			Comprehensive		
	Est.		SE	Est.		SE
Intercept	-0.998	*	0.135	-2.675	*	0.248
Driver 1 age (years)	-0.011	*	0.004	0.039	*	0.008
Driver 1 age squared (hundreds of years)	0.013	*	0.003	-0.048	*	0.007
Driver 1 female	0.067	*	0.021	-0.084	*	0.041
Driver 1 married	0.048		0.025	0.125	*	0.046
Driver 1 separated, divorced, or widowed	0.000		0.023	0.058		0.045
Driver 1 insurance score (tens)	-0.018	*	0.001	-0.013	*	0.001
Has 2 drivers	0.063		0.123	-0.135		0.214
Has 3+ drivers	0.529	*	0.158	0.058		0.255
Young driver	0.020		0.049	0.019		0.082
Driver 2 age (years)	0.012	*	0.005	0.006		0.009
Driver 2 age squared (hundreds of years)	-0.013	*	0.005	-0.002		0.008
Driver 2 female	0.097	*	0.034	-0.064		0.060
Driver 2 married	-0.207	*	0.047	-0.121		0.087
Driver 2 separated, divorced, or widowed	0.088		0.164	0.000		0.302
Vehicle 1 age (years)	-0.012		0.005	-0.028	*	0.006
Vehicle 1 age squared (hundreds of years)	-0.015		0.044	0.143	*	0.036
Vehicle 1 personal use	-0.010		0.014	-0.034		0.025
Vehicle 1 passive restraint	-0.078		0.062	-0.114		0.102
Vehicle 1 anti-theft	0.011		0.015	0.018		0.027
Vehicle 1 anti-lock brakes	0.026		0.016	0.039		0.030
Has 2 vehicles	0.281	*	0.056	0.689	*	0.095
Has 3+ vehicles	0.293	*	0.107	0.930	*	0.156
Vehicle 2 age (years)	-0.023	*	0.003	-0.020	*	0.005
Vehicle 2 age squared (hundreds of years)	0.031	*	0.010	0.019		0.018
Vehicle 2 personal use	-0.019		0.015	-0.035		0.027
Vehicle 2 passive restraint	0.075		0.039	-0.033		0.062
Vehicle 2 anti-theft	0.029		0.018	0.009		0.033
Vehicle 2 anti-lock brakes	-0.003		0.019	-0.023		0.032
Year dummies	Yes			Yes		
Territory codes	Yes			Yes		

* Significant at the 5 percent level.

Notes: Insurance score is based on information contained in credit reports. Territory codes indicate rating territories, which are based on actuarial risk factors, such as traffic and weather patterns, population demographics, wildlife density, and the cost of goods and services.

Table B.6: Regression Parameter Estimates - Home

Tricoverage Sample (294,917 household-years)

	Est.		SE
Intercept	-1.968	*	0.250
Insurance score (tens)	-0.018	*	0.001
Home age (years)	0.003	*	0.001
Home age squared (years)	0.000		0.000
Insured value (tens of thousands of dollars)	0.015	*	0.001
Farm or business	0.098	*	0.047
Primary residence	0.631	*	0.228
Owner occupied	0.121		0.077
Number of families	-0.011		0.007
Masonry construction	0.048		0.029
Distance to fire hydrant (feet)	0.001		0.001
Alarm or other protection	0.019		0.036
Year dummies	Yes		
Territory codes	Yes		
Protection classes	Yes		

* Significant at the 5 percent level.

Notes: Insurance score is based on information contained in credit reports. Territory codes indicate rating territories, which are based on actuarial risk factors, such as traffic and weather patterns, population demographics, wildlife density, and the cost of goods and services. Protection classes guage the effectiveness of local fire protection and building codes.

Table B.7: Association Parameter Estimates - Low and High Insurance Scores

	Insurance score ≤ 744 24,288 households 97,985 obs			Insurance score ≥ 837 18,770 households 100,291 obs		
	Est.		SE	Est.		SE
Variances:						
Collision	0.103	*	0.030	0.118	*	0.047
Comprehensive	0.279	*	0.128	0.508	*	0.183
Home	0.410	*	0.085	0.450	*	0.099
Covariances:						
Collision and Comprehensive	0.132	*	0.027	0.157	*	0.034
Collision and Home	0.064	*	0.016	0.085	*	0.022
Comprehensive and Home	0.210	*	0.036	0.236	*	0.052
Correlations:						
Collision and Comprehensive	0.780	*	0.264	0.642	*	0.221
Collision and Home	0.313	*	0.097	0.369	*	0.127
Comprehensive and Home	0.621	*	0.188	0.493	*	0.150

* Significant at 5 percent level.

Notes: Insurance score at time of first observation. Low and high correspond to bottom third and top third, respectively.

Table B.8: Association Parameter Estimates - Low and High Home Values

	Home value ≤ $120,000 26,007 households 132,117 obs			Home value ≥ $160,000 18,791 households 79,084 obs		
	Est.		SE	Est.		SE
Variances:						
Collision	0.099	*	0.037	0.091	*	0.034
Comprehensive	0.336	*	0.124	0.487	*	0.176
Home	0.438	*	0.060	0.433	*	0.102
Covariances:						
Collision and Comprehensive	0.113	*	0.027	0.151	*	0.032
Collision and Home	0.070	*	0.018	0.090	*	0.018
Comprehensive and Home	0.236	*	0.032	0.228	*	0.048
Correlations:						
Collision and Comprehensive	0.619	*	0.219	0.715	*	0.238
Collision and Home	0.337	*	0.109	0.453	*	0.135
Comprehensive and Home	0.614	*	0.147	0.496	*	0.150

* Significant at 5 percent level.

Notes: Insured value of home at time of first observation, rounded to nearest ten thousand dollars. Low and high correspond to bottom third and top third, respectively.

Table B.9: Association Parameter Estimates - Young and Old Primary Drivers

	Driver 1 age ≤ 50 26,619 households 134,985 obs			Driver 1 age ≥ 60 20,378 households 99,898 obs		
	Est.		SE	Est.		SE
Variances:						
Collision	0.103	*	0.027	0.133	*	0.052
Comprehensive	0.347	*	0.101	0.467		0.293
Home	0.468	*	0.059	0.389	*	0.013
Covariances:						
Collision and Comprehensive	0.126	*	0.023	0.192	*	0.050
Collision and Home	0.063	*	0.016	0.034		0.075
Comprehensive and Home	0.216	*	0.030	0.231	*	0.056
Correlations:						
Collision and Comprehensive	0.670	*	0.181	0.772	*	0.347
Collision and Home	0.289	*	0.083	0.151		0.331
Comprehensive and Home	0.536	*	0.112	0.540	*	0.215

* Significant at 5 percent level.

Notes: Age of driver 1 at time of first observation, rounded to nearest decade. Young and old correspond to bottom third and top third, respectively.

Table B.10: Association Parameter Estimates - Female and Male Drivers

	Driver 1 female 20,699 households 94,536 obs			Driver 1 male 41,726 households 200,381 obs		
	Est.		SE	Est.		SE
Variances:						
Collision	0.123	*	0.052	0.102	*	0.023
Comprehensive	0.321		0.211	0.434	*	0.103
Home	0.448	*	0.069	0.436	*	0.006
Covariances:						
Collision and Comprehensive	0.117	*	0.042	0.136	*	0.020
Collision and Home	0.089	*	0.022	0.053	*	0.027
Comprehensive and Home	0.229	*	0.048	0.230	*	0.028
Correlations:						
Collision and Comprehensive	0.586		0.312	0.647	*	0.144
Collision and Home	0.378	*	0.127	0.251		0.130
Comprehensive and Home	0.604	*	0.240	0.529	*	0.090

* Significant at 5 percent level.

Note: Gender of driver 1 at time of first observation.

Table B.11: Association Parameter Estimates - Married Primary Driver

	Driver 1 married 38, 139 households 188, 270 obs		
	Est.		SE
Variances:			
Collision	0.115	*	0.023
Comprehensive	0.444	*	0.100
Home	0.253	*	0.088
Covariances:			
Collision and Comprehensive	0.127	*	0.020
Collision and Home	0.071	*	0.015
Comprehensive and Home	0.218	*	0.026
Correlations:			
Collision and Comprehensive	0.561	*	0.123
Collision and Home	0.414	*	0.120
Comprehensive and Home	0.650	*	0.156

* Significant at 5 percent level.

Note: Based on marital status of driver 1 at time of first observation.

BIBLIOGRAPHY

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002), *Topics in Modelling of Clustered Data*, Cambridge: Chapman and Hall/CRC.

Barseghyan, L., Molinari, F., Morris, D. S., and Teitelbaum, J. (2012), "Unobserved Heterogeneity in Insurance Claim Rates: Bad Luck or Else?" In Preparation.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. (2011), "The Nature of Risk Preferences: Evidence from Insurance Choices," Available at SSRN: http://ssrn.com/abstract=1646520.

Bishop, C. (2006), *Pattern Recognition and Machine Learing*, New York: Springer.

Cameron, A. and Trivedi, P. (1998), *Regression Analysis of Count Data*, no. 30 in Econometric Society monographs, Cambridge: Cambridge University Press.

— (2009), *Microeconometrics*, Cambridge: Cambridge University Press.

Cohen, A. and Siegelman, P. (2010), "Testing for Adverse Selection in Insurance Markets," *Journal of Risk and Insurance*, 77, 39–84.

Cook, R. and Li, G. Y. (2002), "Marginal Methods for Incomplete Longitudinal Data Arising in Clusters," *Journal of American Statistical Association*, 97, 1071–1080.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Einav, L., Finkelstein, A., and Cullen, M. R. (2010), "Estimating Welfare in Insurance Markets Using Variation in Prices," *Quarterly Journal of Economics*, 125, 877–921.

Fieuws, S. and Verbeke, G. (2006), "Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles," *Biometrics*, 62, 424–431.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Mohlenberghs, G. (2009), *Longitudinal Data Analysis*, Boca Raton: CRC Press.

Gourieroux, C., Monfort, A., and Trognon, A. (1984a), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701–20.

— (1984b), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Guo, X. and Carlin, B. P. (2004), "Separate and Joint Modeling of Longitudinal and Event Time Data using Standard Computer Packages," *The American Statistician*, 58, 1–9.

Hall, P., Ormerod, J., and Wand, M. (2011), "Theory of Gaussian Variational Approximation for a Poisson Mixed Model," *Statistics Sinica*, 21, 369–389.

Heckman, J. and Leemer, E. (2001), *Handbook of Econometrics*, vol. 5.

Jordan, M. (2004), "Graphical Models," *Statistical Science*, 19, 140–155.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233.

Kocherlakota, S. and Kocherlakota, K. (1993), *Bivariate Discrete Distributions*, New York: Marcel Dekker.

Kullback, S. and Leibler, R. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, 35, 13–22.

Lindsay, B. G. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 221–239.

Little, R. and Rubin, D. (2002), *Statistical Analaysis with Missing Data*, New York: Wiley, 2nd ed.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models (Second Edition)*, London: Chapman & Hall.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley Series in Probability and Statistics.

Morris, D. S. (2011), "A Semiparametric Approach for Multivariate Longitudinal Count Data," Job Market Paper.

Opper, M. and Archambeau, C. (2009), "Variational Gaussian Approximation Revisited," *Neural Computation*, 21, 786–792.

Ormerod, J. and Wand, M. (2010), "Explaining Variational Approximation," *The American Statistician*, 64.

— (2011), "Gaussian Variational Approximate Inference for Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, Accepted for Publication.

Pinquet, J. (1998), "Designing Optimal Bonus-Malus Systems from Different Types of Claims," *ASTIN Bulletin*, 28, 205–220.

— (2012), "Experience Rating in Non-Life Insurance," Cahier de recherche 2012-10.

Prentice, R. (1988), "Correlated Binary Regression with Covariates Specific to each Binary Observation," *Biometrics*, 44, 1033–1048.

Rabe-Hesketh, S., Skrondal, A., and Pickels, A. (2005), "Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models with Nested Random Effects," *Journal of Econometrics*, 128, 301–323.

Rizopulos, D., Verbeke, G., and Lesaffre, E. (2009), "Fully Exponential Laplace Approximation for the Joint Modelling of Survival and Longitudinal Data," *Journal of the Royal Statistical Society, B*, 71, 637–654.

Robins, J., Rotnitzy, A., and Zhao, L. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of American Statistical Association*, 90, 106–121.

Tsiatis, A. A. and Davidian, M. (2004), "Joint Modeling of Longitudinal and Time-to-Event Data: An Overview," *Statistica Sinica*, 14, 809–834.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS," *Journal of the American Statistical Association*, 90, 27–37.

Winkelmann, R. (2003), *Econometric Analysis of Count Data*, Berlin: Springer.

Wooldridge, J. M. (2001), *Econometric Analysis of Cross Section and Panel Data*, Cambridge: The MIT Press.

Wu, L., Liu, W., Yi, G., and Huang, Y. (2012), "Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods and Issues," *Journal of Probability and Statistics*, Article ID 640153.

Wulfsohn, M. S. and Tsiatis, A. A. (1997), "A Joint Model for Survival and Longitudinal Data Measured with Error," *Biometrics*, 53, 330–339.